Research Report

Modeling Nonignorable Missing Data With Item Response Theory (IRT)

Norman Rose

Matthias von Davier

Xueli Xu

April 2010

ETS RR-10-11

Modeling Nonignorable Missing Data With Item Response Theory (IRT)

Norman Rose
Friedrich Schiller University, Jena, Germany

Matthias von Davier and Xueli Xu
ETS, Princeton, New Jersey

April 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Dan Eignor

Technical Reviewers: Kentaro Yamamoto, Shelby Haberman

Abstract

Large-scale educational surveys are low-stakes assessments of educational outcomes conducted using nationally representative samples. In these surveys, students do not receive individual scores, and the outcome of the assessment is inconsequential for respondents. The low-stakes nature of these surveys, as well as variations in average performance across countries and other factors such as different testing traditions, are contributing factors to the amount of omitted responses in these assessments. While underlying reasons for omissions are not completely understood, common practice in international assessments is to employ a deterministic treatment of omissions, either as missing observations or as responses that are considered wrong. Both approaches appear problematic. In this project, we analyzed the effects of treating omitted responses either as missing or as wrong, as is done in the majority of international studies, and compared these data-treatment solutions to model-based approaches to treating omitted responses. The two types of model-based approaches used in this study are: (a) extensions of multidimensional item response theory (IRT) with an additional dimension based on response indicator variables defined and calibrated together with the set of items containing the observed responses and (b) multidimensional, multiple-group IRT models with a grouping variable representing the within-country stratification of respondents by the amount of omitted responses. These two model-based approaches were compared on the basis of simulated data and data from about 250,000 students from 30 Organisation for Economic Co-operation and Development (OECD) Member countries participating in an international large-scale assessment.

Key words: multidimensional item response theory, missing data, large-scale assessments, latent regression model, IRT

Table of Contents

List of Tables

Page

List of Figures

Background

Large-scale educational surveys are low-stakes assessments of academic achievement conducted using nationally representative samples. In international surveys such as Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Studies (TIMSS), and Progress in International Reading Literacy Study (PIRLS), more than 50 countries participate in a coordinated administration of translated surveys. The low-stakes nature of these surveys, as well as variations in average performance across countries and other factors such as different testing traditions, have been discussed as contributing factors to the amount of omitted responses observed in these assessments. While it is reasonable to assume that the underlying reasons for omissions are not completely understood, common practice in international assessments is to employ a deterministic treatment of omissions either as missing observations or as responses that are considered wrong. Both approaches appear problematic, since the treatment of omissions as missing data assumes that missingness due to nonresponse is ignorable, and the treatment of omissions as (always) wrong responses assumes that the an omission indicates with certainty that the correct response is unknown, independent of respondent ability.

In the National Assessment of Educational Progress (NAEP), omitted responses to multiple choice items are consistently treated as *partially correct* responses, while omissions on constructed response items are consistently considered to be wrong in psychometric scaling and ability estimation. In some international assessments such as PISA and TIMSS, the missing responses are treated differently in different phases of the analyses. In this project, we analyze the effects of treating omitted responses either as missing or as wrong as is done in the majority of international studies, and compare these data-treatment solutions to the following two model-based approaches in which missingness is treated as nonignorable:

1. Extensions of multidimensional IRT, in our study specified as special cases of a general latent variable framework, the general diagnostic model (GDM; von Davier, 2005). In this approach, an additional dimension based on response indicator variables (Glas & Pimentel, 2008; Holman & Glas, 2005; Moustaki & Knott, 1999) was defined and calibrated together with the set of items containing the observed responses.

2. An extension of the GDM that allows estimation of multidimensional, multiple-group IRT models (Xu & von Davier, 2006, 2008). In this context, a grouping variable was specified that represents the within-country stratification of respondents by the amount of omitted responses. This, in effect, emulates a latent regression model (Mislevy & Sheehan, 1992; von Davier, Sinharay, Oranje, & Beaton, 2007) with omission rate by country as a predictor of average performance

Missing data due to examinees' nonresponse threatens statistical inferences if the target of inference, for example the proficiency variable of interest, and the tendency to omit responses are not independent. Rubin's framework (Rubin, 1976; Rubin & Little, 2002) differentiates among three kinds of missing data according to the underlying missing data mechanism: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). In order to define the three kinds of missing data, Rubin distinguishes between the observed data $\boldsymbol{Y}_{obs}$ and the missing data $\boldsymbol{Y}_{miss}$. Together, these constitute the complete data matrix $\boldsymbol{Y} = (\boldsymbol{Y}_{obs},\boldsymbol{Y}_{miss})$. Here, we adapted this notation to the latent variable framework. $\boldsymbol{Y}$ here is the complete data matrix that consists of the observed item responses $\boldsymbol{Y}_{obs}$ and the omitted responses $\boldsymbol{Y}_{miss}$ of the k items Y_1 to Y_k, indexed by i. The values of a latent variable ξ can also be considered to be missing data. In large-scale assessments, it is a common practice to include covariates such as gender and social economic status in the analyses. All covariates constitute the matrix $\boldsymbol{Z}$. *MCAR* denotes the case where the distribution of missing data is independent of $\boldsymbol{Y}_{obs}$, $\boldsymbol{Y}_{miss}$ and a given multivariate covariate $\boldsymbol{Z}$. That is, $P(\boldsymbol{D} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}, \boldsymbol{Z}, \xi) = P(\boldsymbol{D})$. The matrix $\boldsymbol{D}$ is an indicator matrix consisting of the indicator variables d_i that marks the occurrence of the values of Y_i,

$$d_i = \begin{cases} 1, \text{if } Y_i \text{ is observed} \\ 0, \text{ if } Y_i \text{ is not observed.} \end{cases} \quad (1)$$

MAR holds if the distribution of the missing data is only dependent on the observed variables $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Z}$ but not dependent on the unobserved values of missing data $\boldsymbol{Y}_{miss}$ and ξ. This is equivalent to the expression $P(\boldsymbol{D} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}, \boldsymbol{Z}, \xi) = P(\boldsymbol{D} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Z})$. The third type, called MNAR, can be written as $P(\boldsymbol{D} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}, \boldsymbol{Z}, \xi) \neq P(\boldsymbol{D} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Z})$. It is the opposite of MAR. That means the conditional distribution of the missing data given $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Z}$ depends on the unobserved data $\boldsymbol{Y}_{miss}$ and possibly ξ. MCAR and MAR do not jeopardize likelihood-based or

Bayesian inference. The missing data mechanism can be considered to be ignorable and does not need to be taken into account. However, MNAR implies a violation of the ignorability principle and requires appropriate measures to account for the effects of data MNAR. Different methods have been proposed to account for MNAR. Such methods include regression-based methods, multiple imputation (Rubin, 1987; Schafer, 1997), methods based on the EM algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 1996), as well as weighting procedures (Rubin & Little 2002).

In this paper, the effects of different treatments for nonignorable missing data in the area of educational large-scale survey assessments are considered. In this field, accurate estimates of the item parameters and the proficiencies of reporting subgroups are of major interest. How item and person parameter estimates are affected by missing data is shown by studying measures based on *classical test theory* (CTT) and those derived from models from *item response theory* (IRT). Additionally, it is demonstrated that the occurrence of missing data and its treatment is also a matter of test fairness. For that reason, a case-based simulation was conducted to demonstrate the effects of treating missing data in different ways. Finally, different IRT models are utilized in order to analyze data from PISA, using the database from the 2006 assessment cycle.

Many large-scale assessments are low-stakes surveys, which typically suffer from a substantial amount of missing data. With large-scale surveys, a subset of items, organized into a number of test booklets, is administered to the students, so that a large proportion of all item responses are missing by design. The test booklets are spiraled through the sample so that the assignment of booklets to students is random, and thus, the data missing due to booklet assignment can be viewed as data that is MCAR. In addition to the items that were not administered, additional items may not be responded to by the students, either due to a lack of motivation, a lack of sufficient testing time, or for other reasons. This portion of the missing data is the target of the current study, since these omissions are often related to student characteristics. It has to be assumed that the proportion of missing data may be related to the latent variable of interest, even after conditioning on the observed data and covariates. Using the introduced terminology, that means that the omitted responses are MNAR. To tackle this problem, common practice in operational data analyses tends to replace the omissions in a somewhat ad hoc manner and treats these afterwards as if the data were observed by design in the way chosen by the

analyst. For example, in many large-scale assessments, the missing data are treated as incorrect responses, that is, as if an answer was attempted but the question was *answered not correctly*. That procedure ignores the stochastic relation between the latent proficiency variable and the manifest item response. A deterministic replacement of omitted response by wrong response implicitly assumes that the omitted item can never be solved, regardless of the ability of a person. In some surveys, the treatment of missing data differs between the preliminary stage of item calibration and the later stages of group-level estimation of ability distributions. In the initial item calibration, the missing responses are ignored, that is, treated as MCAR, but in the subsequent operational analyses, the missing data are treated as incorrect responses. The item parameters, however, are taken as fixed, known constants, from the item calibration in the first stage. This, in effect, changes the dataset between item calibration using IRT (where omissions = missing), and ability estimation using latent regression-based population models (where omissions = wrong), and thus changes basic item statistics, such as the percent correct and item total correlations between different stages of the analysis.

Some theoretically derived model-based approaches to dealing with data MNAR have been proposed in the literature. With the *symmetric pattern model*, O'Muircheartaigh and Moustaki (1999) introduced a latent variable approach to tackle the problem of data MNAR. Holman and Glas (2005) as well as Korobko et al. (2008) chose the same idea to account for nonignorable missing data. These authors used the matrix $\boldsymbol{D}$ of the indicator variables d_i (see Equation 1) to establish a measurement model for a so-called latent response propensity θ. The person estimators of θ can be used to compute weights for each observation (Moustaki & Knott, 2000). Alternatively, the *measurement model* for the response propensity θ has been added to the measurement model for the ability variable of interest, ξ. This results in a multi-dimensional IRT (MIRT) model and the information on the missing data with respect to ξ is directly taken into account in the maximum likelihood estimation. The MIRT model can be specified in different ways to account for the relationship between ability and response propensity. MIRT models with either a between-item-multidimensional structure or a within-item-multidimensional structure have been introduced to account for nonignorable missing data. Another possibility is to avoid modeling a second latent variable by regarding the missing value code as an additional response category in a nominal response model (Moustaki & Knott, 2000). The basic idea of these models involves the specification of a selection model (Heckman, 1979) that accounts for the stochastic

dependency between the latent proficiency ξ and the occurrence of missing data. In addition to the above approaches, we propose taking the missing data mechanism into account by specifying a predictor variable of the latent proficiency ξ to be introduced in a latent regression model. The proposed approach has, in our view, at least three advantages:

1. The latent regression approach does not increase the dimensionality of the IRT scaling model by an additional response propensity dimension based on the indicator variables d_i,
2. A covariate based on the amount of missingness can be defined even in cases where there are too few omissions to support IRT-based scaling of response propensity indicator variables (when d_i are all very easy, i.e., the amount of omissions is low),
3. The latent regression approach comes at virtually no cost for large-scale survey analyses such as the ones conducted in PISA, TIMSS, and NAEP, since these involve estimation of latent regression models with a large number of predictors based on student background data. A response propensity indicator would simply act as an additional predictor in these models.

In subsequent sections of this paper, we will compare these model-based approaches to the operationally used approaches that treat the omitted responses in deterministic ways. For these latter deterministic approaches, the observed item response data matrix is replaced by a new matrix, in which omitted responses are either replaced by wrong responses or by codes that indicate responses are missing at random. Little attention has been given to the effects of different missing data replacements of this type of item and person statistics commonly used in CTT. The sum score of a person over all items or functions of that score are often used as a simple estimate of ability, and are obviously affected by the way one treats omitted responses. In the presence of missing data, the mean overall completed items of a person can be used. A measure of the difficulty of the item Y_i in CTT is the expectation $E(Y_i)$ estimated by the sample mean of that item. In a case-based simulation study described in this paper, the effects of nonignorable missing data and its treatment will be demonstrated with respect to these measures and to the IRT parameters of different models. These IRT models will be introduced next.

Item Response Theory Models

In the case-based simulation study as well as in the analysis of the real data from PISA 2006 described in this paper, we compared IRT models that ignore the missing data, as well as IRT models that take the information about nonignorable missing data into account in model-based, nondeterministic ways. This was done in order to compare the performance of the common practices of handling missing data in large-scale assessments with model-based approaches to dealing with nonignorable missing data.

Altogether, seven different models were considered. Some models were multidimensional generalizations of the Rasch model; some were generalizations of the two-parameter logistic IRT (2PL IRT; Lord & Novick, 1968) model. The generalizations of the Rasch model used in this study are special cases of the mixed-coefficients multinomial logit model (MCMLM; Adams & Wu, 2007) and can be specified in the Conquest software (Wu, Adams, Wilson, & Haldane, 2007). Both the generalizations of the Rasch models used in this study and the generalizations of the 2PL model used in this study are special cases of the general diagnostic model (GDM; von Davier, 2005), and can be estimated with the mdltm software (von Davier, 2005). We will indicate below which software was used in each case. Not all the models were used in both the simulation study and the PISA real-data analysis. The differences between the models used in this study are described in the following:

- Model 1 is the unidimensional IRT model using the complete data Y. The IRT model for the complete data was only used in the simulation study, since complete data is unavailable with real datasets. Therefore, Model 1 was estimated only in the simulation case-study in order to provide a gold standard for comparison.
- Model 2 is the unidimensional IRT model where the missing data were ignored. This model assumes implicitly that the missing data are missing completely at random. The model specification is equivalent to Model 1.
- Model 3 is the unidimensional IRT model using data where omitted responses were treated as *incorrect*. The model specification is equivalent to Model 1 and 2 but estimation is based on recoded data.
- Model 4 has also the same model specification as Models 1 to 3, but the item parameters were taken as fixed from Model 2. This model emulates the procedures in

many large-scale assessments. That is, in the item calibration stage the missing data are ignored. In the subsequent analysis, where the person parameter and its distribution are estimated given the fixed item parameters from the item calibration, the missing data are treated as *wrong*.

The Models 2 to 4 were also applied to the real PISA 2006 data. Because of the complexity of these data, some extensions had to be introduced. First, the latent ability variable needed to be multidimensional. Therefore, four latent variables for mathematics, reading, science and the latent response propensity were incorporated. Second, the country variable was included in the model as a covariate, by specifying a multiple-group model. The item parameters were constrained to be equal across the 30 Organisation for Economic Co-operation and Development (OECD) countries used in this study. In order to compare the countries with respect to the distributions of the latent variables, their means, variances, and covariance terms were freely estimated within each country.

Model 5 is the latent regression model, where ξ is regressed on the observed response rate $\bar{d}(U)$ of a person U. We assume a general tendency of examinees to complete items. This response propensity is also a latent variable, denoted by θ. The observed response rate of a person is a fallible measure of θ. Figure 1 shows the conceptual path diagram of the latent regression model.

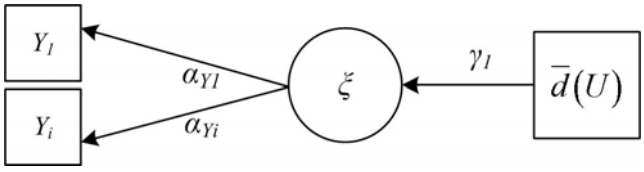

Figure 1. **Conceptual path diagram of the latent regression model.**

Considering the likelihood of Model 5, it can be seen that the missing data information is taken into account in the estimation of the density $g(\xi_u)$ of the latent variable (Equations 1 and 2),

$$L = \prod_{u=1}^{N} \prod_{i=1}^{k} P\left(Y_i \mid \xi_u, \alpha_{Yi}, \beta_{Yi}, d_i\right) g\left(\xi_u\right). \tag{2}$$

Model 5 was applied to the PISA 2006 data in a slightly altered fashion. Instead of using a linear regression, $E\left[\xi \mid \bar{d}\left(U\right)\right]$, the observed response rate was stratified into a number

of groups. Three strata were used (low, medium and high response rates). The distributions of the response rates in the three strata averaged across the countries are given in Table 1, in terms of the mean and the standard deviation. The categorization criterion used was able to obtain nearly equal-sized groups.

Table 1

Sample Size n_{strata}, of the Established Strata With Respect to the Estimated Response Propensity, and Mean and Standard Deviation of the Response Rates Within the Three Strata

Strata	n_{strata}	Mean resp. rate	SD
Mean response rate < 0.90	84,473	0.765	0.129
0.90 ≤ Mean response rate ≤ 0.98	76,797	0.944	0.022
Mean response rate > 0.98	89,872	0.995	0.009

The formed groups for the analysis were all the combinations of the stratified observed response rate and the country variable. Actually, 90 groups resulted. However, due to very low rates of omissions, the data from the Netherlands were classified into only two strata. Therefore, Strata 1 and Strata 2 were combined for the Netherlands and an 89-group model was analyzed as the final model. This model is equivalent to a multiple group latent regression IRT model, with the dummy-coded response strata variable as a predictor (Mislevy & Sheehan, 1992; von Davier, Sinharay, Oranje, & Beaton, 2007).

- Model 6 is the between-item-multidimensional model with two latent variables ξ and θ. The measurement model for θ consists of the response indicator variables d_i. Consequently, this model is more complex because it needs two latent variables and double the number of items when compared to the unidimensional models and the latent regression model. The left panel of Figure 2 shows a path diagram illustrating the model structure.

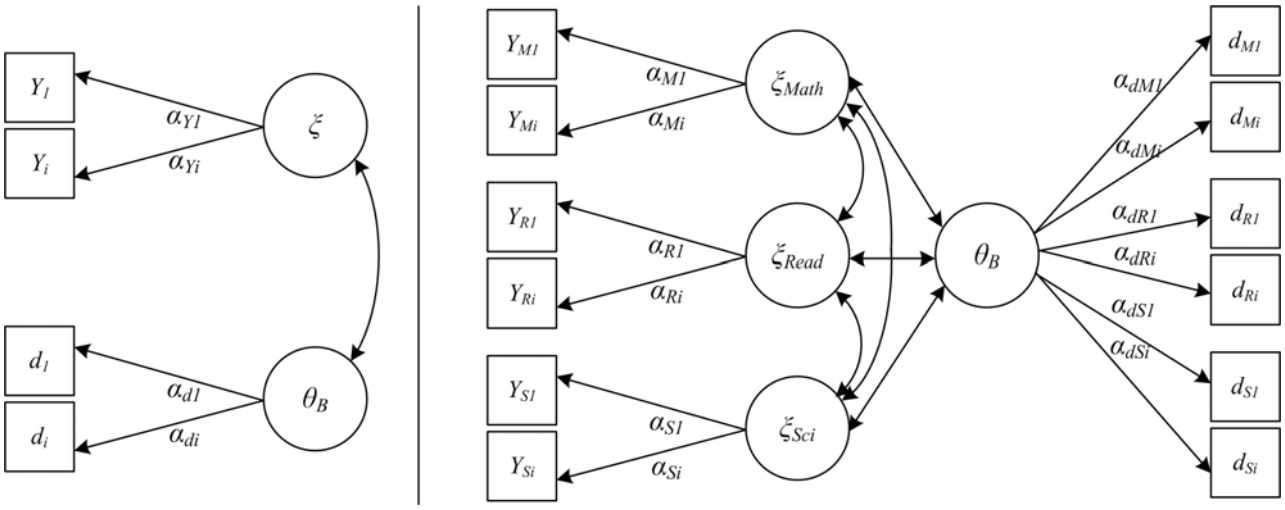

Figure 2. **Conceptual path diagram of the between-item-multidimensional model (left) and the between-item-multidimensional model that was estimated for the PISA 2006 data (right).**

A closer look at the likelihood of Model 6 (Equation 3), reveals that the probability $P(Y_i = 1 | \xi, \alpha_{Yi}, \beta_{Yi})$ is weighted by the $P(d_i = 1 | \theta_B, \alpha_{di}, \beta_{di})$, the response propensity to item Y_i, during the estimation. That corrects the parameter estimates for nonignorable missing data

$$L = \prod_{u=1}^{N} \prod_{i=1}^{k} P\left(Y_i \mid \xi_u, \alpha_{Yi}, \beta_{Yi}, d_i\right) P\left(d_i \mid \theta_{Bu}, \alpha_{di}, \beta_{di}\right) g\left(\xi_u, \theta_{Bu}\right). \qquad (3)$$

We denote θ as θ_B to distinguish this variable from θ_W in Model 7. The variable $g(\xi_u, \theta_{Bu})$ is the joint distribution of the latent ability variable ξ and the latent response propensity θ_B. The model allows for the estimation of the covariance $\sigma(\xi_u, \theta_{Bu})$ of the two latent variables, which is a measure that in some sense quantifies the amount of nonignorability of the missing data.

In order to apply Model 6 to the data of PISA 2006, some extensions had to be introduced to account for the complexity of the data. First, the latent ability variable needed to be multidimensional. Therefore, four latent variables for mathematics, reading, science, and the latent response propensity were incorporated. Second, the country variable was included in the model, by specifying a multiple group model. The item parameters were constrained to be equal across the countries. The variances, covariances and means of the latent variables were freely

estimated for each country. Figure 2 (right) shows the path diagram of the model for each country.

Model 7, the last model considered in this study, is the within-item-multidimensional IRT model, with θ_W as a latent variable that contains the missing data information. As Figure 3 reveals, the measurement model contains additional item discriminations $\alpha_{di;\xi}$, which relates the response indicators d_i with the latent ability ξ.

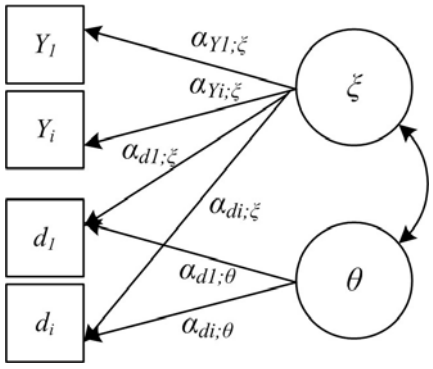

Figure 3. **Conceptual path diagram of the within-item-multidimensional model.**

The likelihood of this model is equivalent to Equation 3, but the term $P(d_i = 1 | \theta_B, \alpha_{di}, \beta_{di})$ is replaced by $P(d_i = 1 | \xi, \theta_B, \boldsymbol{\alpha}_{di}, \beta_{di})$. Model 7 is more complex in that it allows multiple loadings per item for the response propensity variables d_i. To identify Model 7, restrictions have to be set. Depending on the chosen restrictions the meaning of θ_W varies as well as the associated uni- and bivariate statistics (e. g., $\mu(\theta_W)$, $\sigma(\xi, \theta_W)$). In the case of the Rasch model and under particular restrictions in the 2PL model, Model 7 can be transformed into Model 6, and vice-versa. Model 6 is much easier to interpret. For that reason, Model 7 was not applied to the PISA 2006 data.

With the 2PL model, which was used in the PISA 2006 analysis, the general model equations of the items Y_i in the measurement model for ξ are equal in all seven models. It is given with Equation 4.

$$P\left(Y_i \mid \xi, \alpha_{Yi}, \beta_{Yi}\right) = \frac{\exp\left[\alpha_{Yi}\left(\xi - \beta_{Yi}\right)\right]}{1 + \exp\left[\alpha_{Yi}\left(\xi - \beta_{Yi}\right)\right]}. \qquad (4)$$

The model equation for θ depends on the type of multidimensionality. For the between-item-multidimensional model, Model 6, it is given by Equation 5, whereas Equation 6 is the

model equation for the response indicator variables in the within-item-multidimensional model, Model *7*.

$$P\left(d_i \mid \theta_B, \alpha_{di}, \beta_{di}\right) = \frac{\exp\left[\alpha_{di}\left(\theta_B - \beta_{di}\right)\right]}{1+\exp\left[\alpha_{di}\left(\theta_B - \beta_{di}\right)\right]}. \tag{5}$$

$$P\left(d_i \mid \theta_W, \boldsymbol{\alpha_{di}}, \beta_{di}\right) = \frac{\exp\left(\alpha_{di;\xi}\xi + \alpha_{di;\theta}\theta_W + \beta_{di}\right)}{1+\exp\left(\alpha_{di;\xi}\xi + \alpha_{di;\theta}\theta_W + \beta_{di}\right)}. \tag{6}$$

While Models 1 to 4 do not account for the stochastic nature of the missing data, by ignoring or by recoding the missing data, Models 5 to 7 are more complex and model the missing data mechanism alongside the measurement model of the latent ability ξ. We expect that the complexity is rewarded by more accurate, more reliable, and more valid results.

Software

For simplicity, in the simulation case studies only the one-parameter logistic (1PL) Rasch model was utilized. ConQuest (Wu et al., 2007) was used to analyze the simulated data sets. As just mentioned, for the PISA 2006 data we decided to apply the 2PL model. The mdltm software (von Davier, 2005) was used for calibration purposes with this model. This software allows for the estimation of one- and two-parameter logistic versions of general cognitive diagnostic models (von Davier, 2005; von Davier, DiBello, & Yamamoto, 2006; Xu & von Davier, 2006, 2008). The model equation of the GDM is given by

$$P\left(Y_i = y \mid \boldsymbol{\beta}_i, \boldsymbol{\alpha}_i, \mathbf{q}_i, \boldsymbol{\xi}\right) = \frac{\exp\left[\beta_{yi} + \sum_{k=1}^{K} y\alpha_{ik} q_{ik} \xi_k\right]}{1+\sum_{c=1}^{C}\exp\left[\beta_{ci} + \sum_{k=1}^{K} c\alpha_{ik} q_{ik} \xi_k\right]}. \tag{7}$$

The model equation describes the probability of a response in category y of item Y_i given the model parameters. This is a multidimensional model with K latent skills ξ_k. The parameters q_{ik} are elements of the design matrix Q. The entries in this matrix are quantities that relate the latent skills ξ_k to items Y_i. They are not estimated but are part of the model specification. A value $q_{ik} = 1$ means that the latent variable ξ_k influences the category probabilities of item Y_i, whereas $q_{ik} = 0$ means the item response on Y_i is conditionally stochastic independent of ξ_k given the remaining latent variables ξ_t with $t \neq k$ in the model. The parameters α_{ik} denotes the item discrimination

parameters and β_{yi} are the category specific threshold parameters. The crucial difference between diagnostic models and the usual multidimensional IRT models is the discrete nature of the latent skill variables ξ_k. However, by specifying a sufficient number of latent skill levels, the model is equivalent to the 2PL multidimensional model for dichotomous items Y_i, and the generalized partial credit model for ordered categorical items Y_i (Haberman, von Davier, & Lee, 2008).

Simulation Case Study

To study the effects of missing data, a simulation with $N = 1{,}000$ cases and 26 dichotomous items Y_i was conducted. To keep the examples simple, the Rasch model was chosen. The R program was utilized to generate the data. The parameter estimation was carried out using ConQuest.

The parameters for the simulation study were chosen in order to emulate some of the properties of the real PISA 2006 data. However, some of the properties of the real data were altered to be more extreme, to provide a better illustration of effects. For example, the overall proportion of missing data was increased to 30%. Given item Y_i, the proportion of missing data depends on the item difficulty. The easier the item, the greater the expected number of simulated examinees who complete the item. The resulting dependency between the observed item means $\overline{Y}_i(obs)$ of the items Y_i and the mean $\overline{d}(Y_i)$ of the respective response indicators d_i is depicted in Figure 4. The mean $\overline{d}(Y_i)$ is simply the mean of the ith column of the $\boldsymbol{D}$ matrix and is the proportion of persons that completed the item Y_i, regardless of whether they answered correctly or not. The means $\overline{Y}_i(obs)$ of dichotomous items is the proportion of correct answers on all observed item responses to item Y_i. The correlation for the simulated data is $r\left[\overline{Y}_i(obs), \overline{d}(Y_i)\right] = 0.622$, while in the PISA 2006 data, this correlation was only about 0.33. Figure 4 shows the relationship graphically. That dependency was accomplished by establishing a systematic relationship between the threshold parameters of the measurement model for ξ and θ (Figure 4, left).

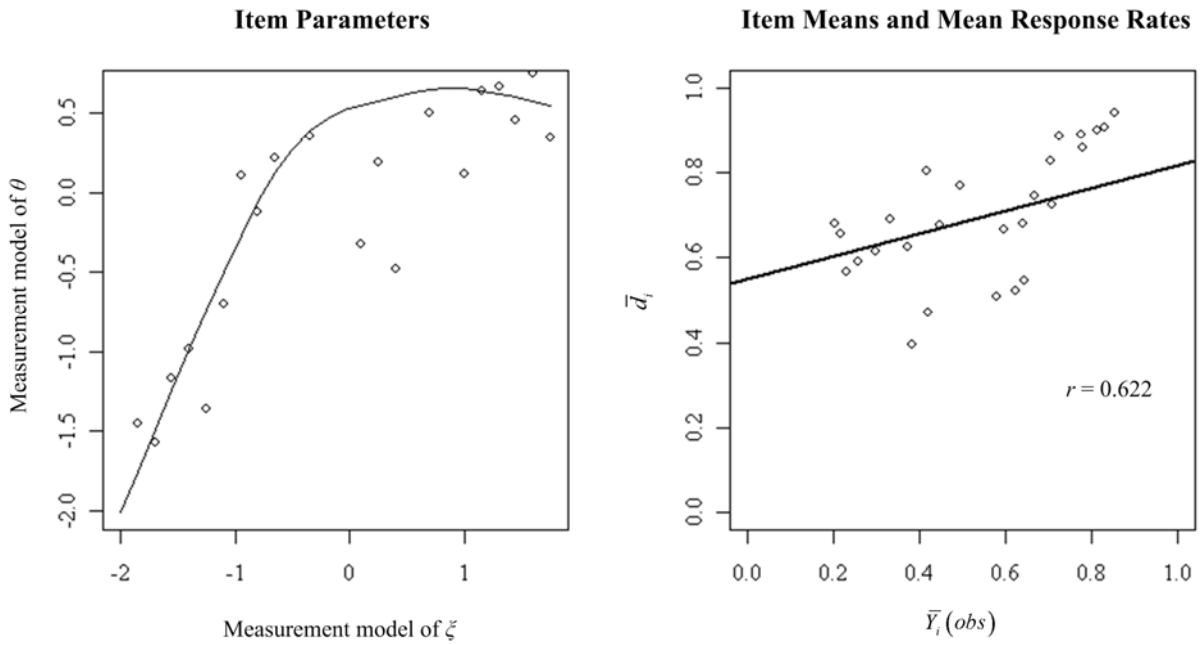

Figure 4. **Item parameters of the both measurement models of ξ and θ (left). The right figure shows the resulting relationship between the item means of the observed item Y_i that indicates ξ and the respective response indicators d_i.**

Under these chosen simulation conditions, the easy items contain fewer missing responses than the difficult items and are answered by almost all examinees. The more difficult an item, the fewer the expected number of persons who answer the item. This dropout is systematic, in the sense that people with lower ability levels tend not to provide a response. That means that with increasing item difficulty, the item mean is estimated on the basis of a less and less representative subsample, with respect to the latent proficiency variable. The observed data consists more and more of responses from more highly proficient people. Consequently, the item means $\overline{Y}_i(obs)$ will be biased estimators that systematically overestimate the true expectation *E(Yi),* due to the selection process. The resulting bias appears to increase with the difficulty of the item. Figure 5 (left) depicts the selection process for the simulated data. The conditional distributions of ξ given the Item Y_i is completed are shown. The means $\overline{\xi}|_{d_i=1}$ of these distributions are correlated with the item difficulty by $r = 0.704$. In CTT, the item parameters are specific to the population employed. For that reason, each item difficulty is specific to a different

subpopulation characterized by its conditional distribution $\xi|_{d_i=1}$. In general, it holds that if the covariance between ρ(ξ, θ) is positive and the probability to respond to the item P(d_i = 1| θ) < 1, then $E\left(Y_{i;obs}\right) \geq E\left(Y_{i;comp}\right)$. Therefore, the expectation of item Y_i is higher for the observed data than for the complete data. The proof is given in Appendix B. This was also found in the simulated data (Figure 5, left). The extent of the bias is mainly driven by two factors: (a) the proportion of missing data and (b) the strength of the stochastic dependency between ξ and the occurrence of missing data. The latter corresponds to the strength of the nonignorability of the missing data mechanism.

Likewise, recoding the missing data into answered incorrectly leads to a systematically biased estimator of the true item mean. Since it is implicitly assumed that the not observed values $Y_{i;miss}$ are zero for all cases, the more the missing data are present, the heavier the bias. Figure 5 (right) shows the realized bias in the simulated data.

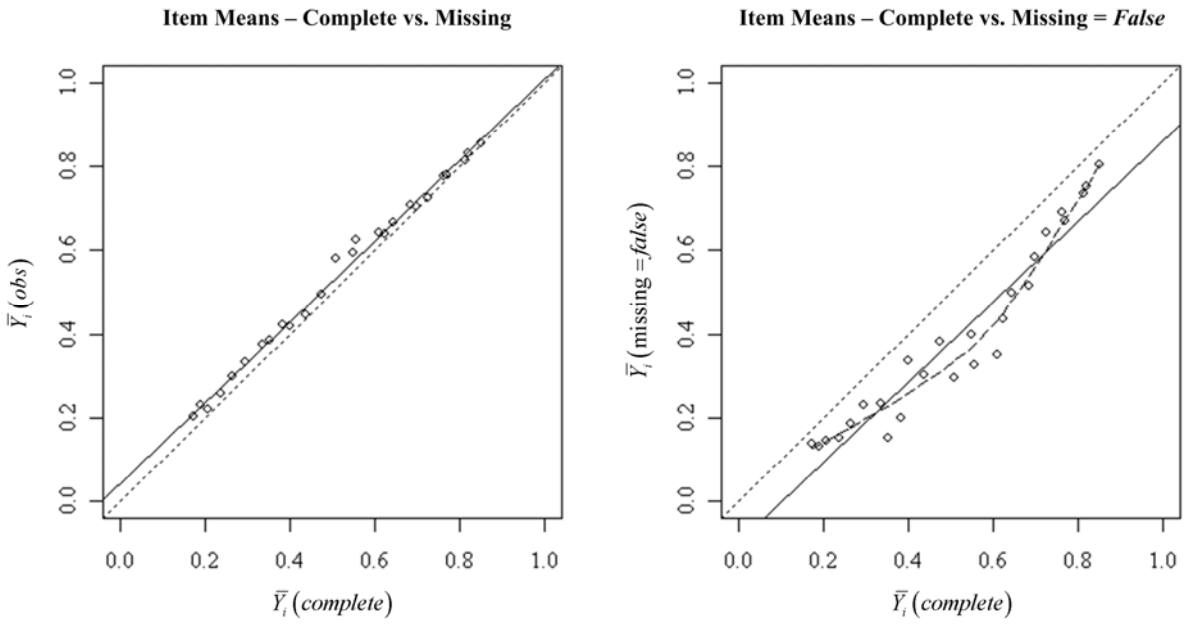

Figure 5. **Comparison of the item means between the complete data and the data with missing responses (left) and the missing data recoded as answered *false*. The solid line shows the linear relation between the parameters. The dashed line (right) represents the smoothing spline and the dotted line is the bisectric.**

So far, only the item measures have been considered. What are the consequences of the selection process in terms of the person estimates? As mentioned above, the sum score or functions of the sum score are used as ability estimates in CTT. In the presence of missing data, the sum score is not useful, because it is implicitly assumed that $Y_{i;miss} = 0$. A more adequate route is to use the mean of all completed items. In the cases of dichotomous items Y_i, this is simply the proportion of correctly answered items on all completed items of person u, denoted as $P^+(u)$. But how is this measure influenced if the missing data mechanism is nonignorable? If the ability ξ and the latent response propensity are positively related and, simultaneously, the item difficulties are related to the proportion of missing of the items, then on average, respondents should choose to answer those items that are easy enough for their ability. That means respondents should tend to complete those items that have a higher probability to be solved, given their value on ξ. That leads to a systematic bias in the proportion correct measures. Respondents with a lot of missing data have $P^+(u)$ that are based on a very different, self-selected, test form. In other words, the measures are not comparable any more if the number of completed items differs systematically. Figure 6 (left) shows the means $\overline{\beta}(u)$ of the item difficulties of the items that person u, with a certain value of ξ, has completed. From a point of view that assumes a rational response process, the missing data mechanism is plausible under the assumption that respondents have a heuristic to judge or develop a vague idea about the difficulty of an item. As a result, they should more likely omit items they judged as too difficult. Therefore, more benefit can be expected in terms of higher proportions of correctly answered items in the data with missing data in the middle and lower range of the ability distribution. Another explanation is that items are not reached, because the speed of processing the items is slower for respondents with lower ability ξ. In that case, the proportion of missing data would also be correlated with ability, but the selection process for the items would be quite different. Respondents would not choose items due to item difficulty but the later items are more often completed, on average, by the more proficient respondents. In this case, the benefit from having missing data depends on the difficulty of the last items compared to the first ones.

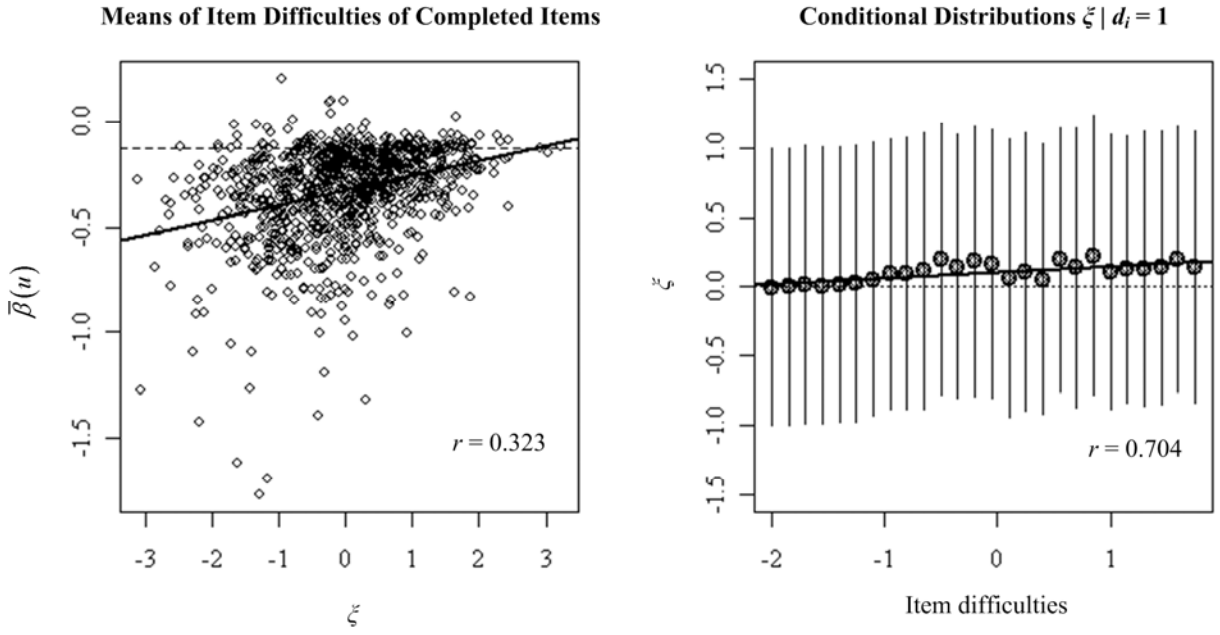

Figure 6. **Left: Mean of the item difficulties averaged across the completed items for each person. The dashed line marks the unconditional mean of the item difficulties. Right: Means of the conditional distributions of ξ given the items were solved. The circles mark the means of the conditional distributions, the vertical black lines indicates +/- 1 SD.**

In our simulation case studies, we emulated the process of an individual judging items and choosing the easier items to complete. If we compare the proportion correct between the complete data (denoted as $P^{+}_{comp}(u)$) and the observed data (denoted as $P^{+}_{obs}(u)$), respondents with missing data should benefit from the missing data mechanism. Figure 7 (left) confirms this supposition. Of course, the process is stochastic and for some respondents the missing mechanism will be disadvantageous.

In the following section, the effects of recoding omitted responses to answered incorrectly will be presented. The proportion of correct answers after recoding $P^{+}_{m=false}(u)$ can only be equal or smaller than the proportion correct $P^{+}_{comp}(u)$ of the complete data. Therefore, handling of missing data in this way constitutes a penalization. Figure 7 (right) reveals that the penalization varies across the distribution of ξ. It is, on average, the highest in the middle to higher range of the latent variable ξ, although the highest rate of omitted responses are observed in the lower

range of ξ. In the lower ability range, more missing data occurred but the skipped items were more likely not answered correctly.

These findings indicate that the occurrence and the treatment of missing data is also a matter of test fairness. A comparison among respondents and among groups of respondents might be unfair if they differ in their amount of missing data and in the strength of the relationship between the latent variable ξ and the missing data. Recoding missing values into incorrect responses might seem plausible at first glance, but the results here indicate that this leads to unfair penalization and wrong inferences. Using this strategy leads to the treatment of missing data not as what they actually are, not observed responses. Quite the contrary, by recoding the missing data into incorrect responses, the dataset is treated as if no missing data had occurred. As mentioned above, within the IRT framework, several model-based approaches have been proposed to handle nonignorable missing data. In the next section, the simulated data were used to study the effect of missing data on the item and person parameter estimates if the missing data are ignored or recoded as wrong. Next, the models that account for the missing data are considered.

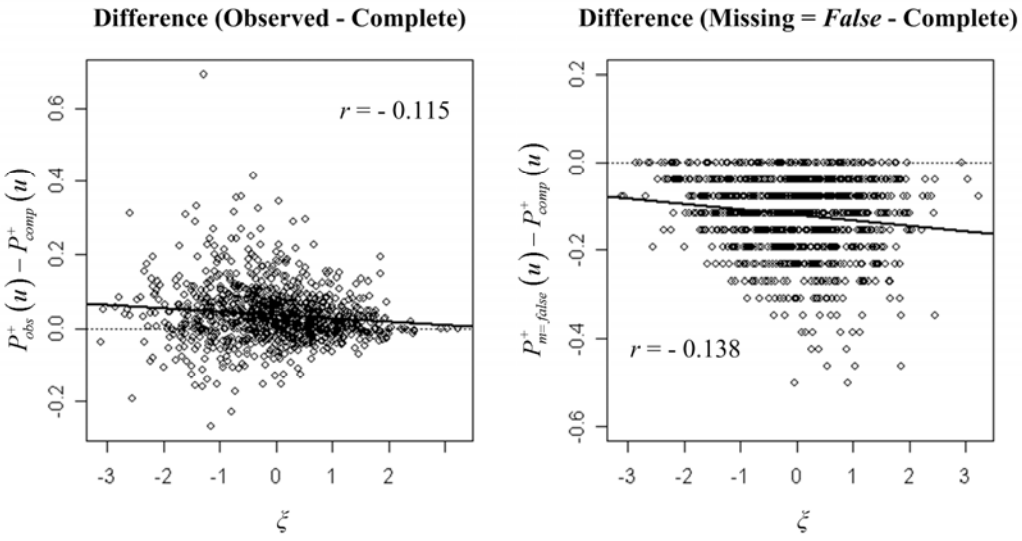

Figure 7. The **difference between the proportion correct of the observed and the complete data (left) and the difference between the proportion correct if the missing data are recoded to false and the complete data (right).**

IRT Models Used in the Simulation Case Study

The Rasch model was chosen for the simulation. Consequently, all item discriminations are restricted to be one. The parameter estimation for the seven IRT models introduced above was carried out by specifying a Rasch model with ConQuest.

Figure 8 displays the results for the item parameters of the measurement model of ξ.

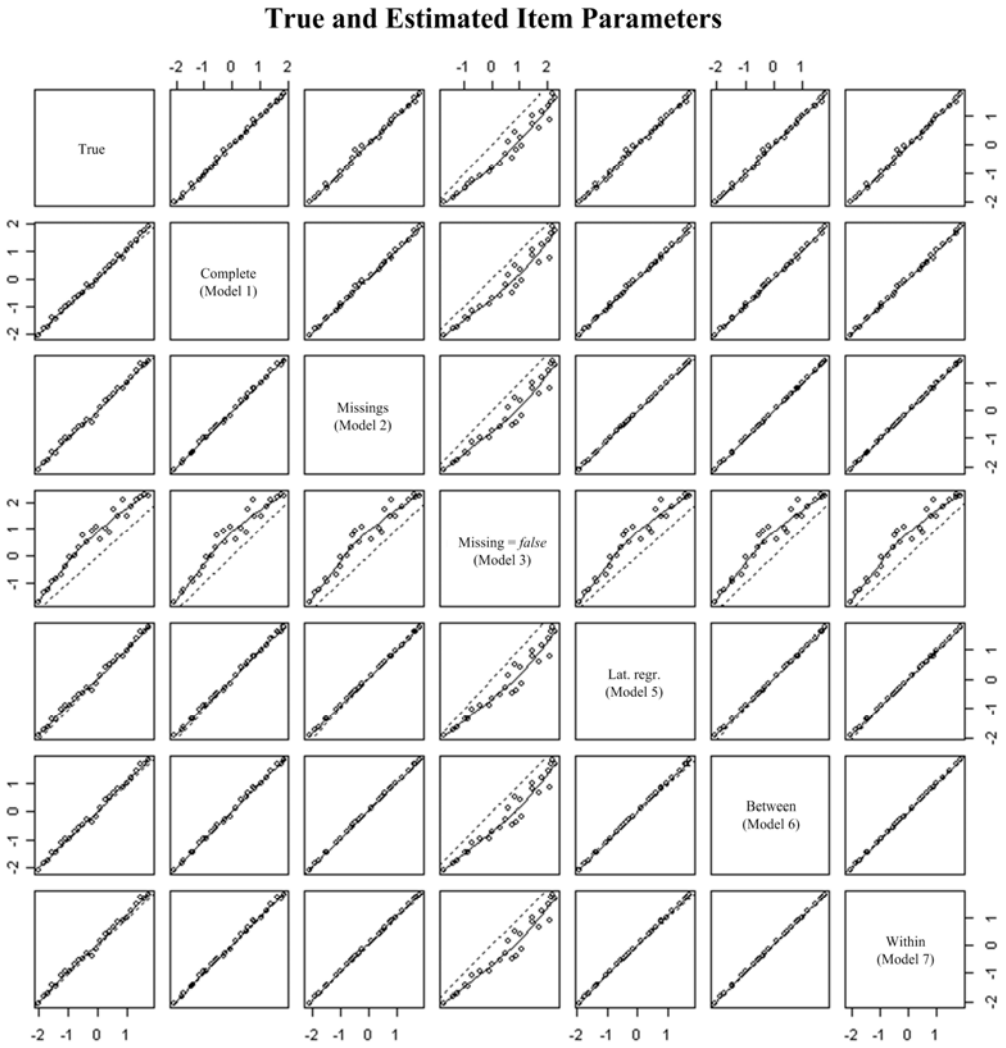

Figure 8. **Pair-wise comparisons of estimated item difficulties: The first row/column illustrates the accuracy of the estimates of the different models by the comparison with the true items difficulties. The dotted line is the bisectric. The solid lines represent smoothing splines.**

Compared with the true item parameters, the estimates of all models fit pretty well, with the exception of Model 2 where the missing data are recoded to *false*. Even Model 3 without modeling the missing data mechanism alongside the measurement model for ξ gives reasonable accurate results with respect to the recovery of item parameters.

Table 2

Bias, Standard Error,s and Mean Squared Error of the Item Difficulties of the Measurement Model of the Latent Proficiency Variable ξ

Model	Item parameter estimates		
	Bias	SE	MSE
Model 1 - complete data	0.0457	0.006	0.00212
Model 2 - ignoring missing data	0.007	0.00961	0.00014
Model 3 - missing = false	0.7362	0.07561	0.54766
Model 5 - latent regression	0.05713	0.00998	0.00336
Model 6 - between-item-multidimensionality	0.0492	0.00913	0.00251
Model 7 - within-item-multidimensionality	0.0649	0.009	0.00429

It is important to note that a single simulation trial is far from sufficient to make a decision about which model to use. The simulation in this paper serves only to illustrate the influence of missing data treatments on parameter estimates. Therefore, a general superiority of Model 2 cannot be concluded, even if in this case, it seems to provide the smallest bias. Note that in this particular case, Model 1 that uses the actual complete data does not fit as well as Model 2 does.

Models 6 and 7 produce virtually identical estimates of item parameters for the measurement model of the latent ability ξ. This result can be explained by the fact that in the special case of the Rasch model (and a few constrained versions of the 2PL), these two

models are identical, up to parameterization. More specifically, this means that the parameters of one model variant can be converted into the other model's parameterization by means of a simple transformation. Appendix A explains this equivalency in more detail. Yung, Thissen, and McLeod (1999), as well as Rijmen (2009) provided similar results for related models such as the bifactor IRT models (Gibbons & Hedeker, 1992).

To summarize this section, the recovery of item parameters is adequate across 6 of the 7 models, with the exception being Model 3. The treatment of missing data as answered not correctly leads to heavily biased item parameter estimates. As expected, they are consistently overestimated. Using the simple unidimensional Rasch model and ignoring the missing data mechanism is quite robust when compared to CTT-based item measures. Three models that account for nonignorable missing data were used, the between- and the within-item multidimensional IRT (MIRT) models, as well as a latent regression model with stratified response propensity groups. The choice between the two MIRT models does not depend on the assumption about the missing data mechanism. Both models are equally suited to account for nonignorable missing data. The crucial difference between the models is the meaning of the two latent variables θ_W and θ_B. The advantage of Model 6 is the clear interpretability of θ_B as a latent propensity score. Accordingly, the correlation $\rho(\theta_B,\xi)$ is a direct measure of the relationship between the latent ability of the observational units and their general tendency to respond to the items. Apart from that, there are no differences with respect to the item parameters in the measurement model of ξ between Model 6 and 7. The remaining question is, how accurate can the variable ξ be estimated, using the different models? That question will be considered in the next section.

To compare the models according to the person parameter estimates, expected a posteriori estimates were used. In the estimation procedure for these parameters, the joint distribution $f(\xi,\theta \mid \mathbf{\Sigma})$ of the latent variables is taken into account. That corrects the estimate of ξ for the missing data.

$$\begin{aligned}\hat{\xi}_u^{EAP} &= \int_\xi \int_\theta \xi \cdot f\left(\xi,\theta \mid \boldsymbol{y}_{obs}^{(u)},\boldsymbol{d}^{(u)}\right) d\xi d\theta \\ &= \frac{\int_\xi \int_\theta \xi \cdot P\left(\boldsymbol{y}^{(u)} \mid \xi,\boldsymbol{d}^{(u)}\right) P\left(\boldsymbol{d}^{(u)} \mid \theta\right) f\left(\xi,\theta \mid \Sigma\right) d\xi d\theta}{P\left(\boldsymbol{y}^{(u)} \mid \boldsymbol{d}^{(u)}\right)}.\end{aligned} \tag{8}$$

The variable $\hat{\xi}_u^{EAP}$ is the EAP estimate for person u and $\boldsymbol{\Sigma}$ denotes the variance-covariance matrix of ξ and θ. The accuracy of the EAPs of the seven IRT models are summarized in Table 3 and Figure 9.

Table 3

Bias, Standard Errors, and Mean Squared Error of the EAP Estimators of the Latent Proficiency Variable ξ

Model	Person parameter estimates (EAP)				
	Bias	SE	MSE	$r\left(\hat{\xi}_{EAP}, \xi\right)^2$	$\text{Rel}\left(\hat{\xi}_{EAP}\right)$
Model 1—Complete data	0.03623	0.43499	0.19053	0.814	0.822
Model 2—Missing = 9	0.03535	0.50994	0.26129	0.744	0.757
Model 3—Missing = false	0.03338	0.53353	0.28577	0.723	0.830
Model 4—Missing = false (2)	-0.62029	0.55965	0.69797	0.720	0.756
Model 5—Latent regression	0.03530	0.48290	0.23444	0.770	0.782
Model 6—Between-item-multidimensionality	0.03562	0.48211	0.2337	0.771	0.781
Model 7—Within-item-multidimensionality	0.04108	0.48236	0.23436	0.771	0.791

Note. $\text{Rel}\left(\hat{\xi}_{EAP}\right)$ denotes the EAP-reliability, defined as the ratio of the variance of the EAPs and the variance of the plausible values.

With the exception of Model 4, the bias does not differ substantially across models. Even Model 2 and 3 have small biases. Treating the missing data as answered not correctly, while using the item parameter estimates from Model 1, leads to heavily biased person parameters for Model 4. As expected, the EAPs from Model 4 underestimate the true ability. A comparison of the estimates obtained with Models 3 and 4 shows that these models are indeed only different parameterizations of the same underlying structural assumptions. For the other models, only small differences can be seen with respect to the standard errors and the MSE of the estimates. According to the considered measures, Models 5 to 7 seems to be slightly better

than Models 2 to 4. The results suggest that accounting for nonignorable missing data by modeling the missing data mechanism as is done in Models 5 to 7 improves the accuracy of the EAP estimates.

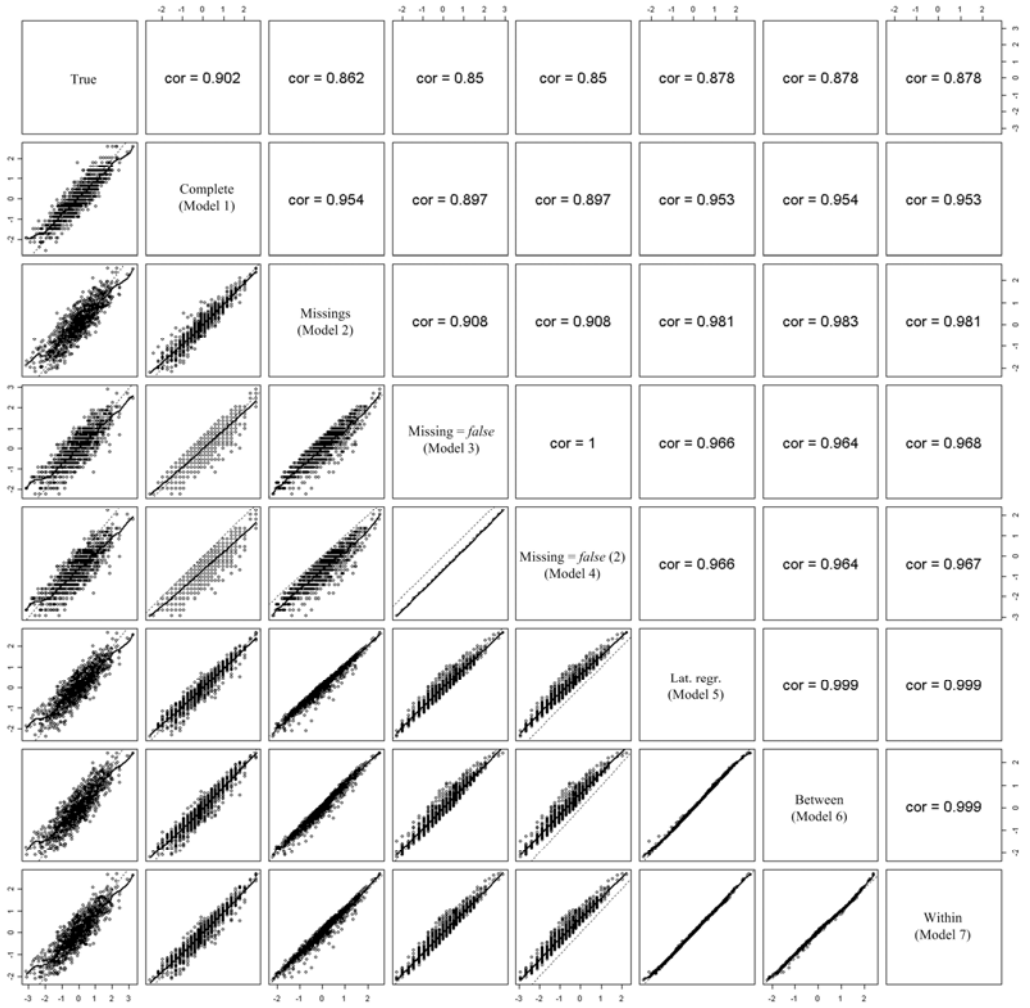

Figure 9. **EAP parameter estimates of ξ. The true person parameters of ξ used to simulate the data is indicated by *true*. The correlations are depicted in the upper triangle. The dotted line is the bisectric and the solid line is a smoothing spline regression.**

In addition, Figure 9 provides the correlation between the true values of ξ and the different estimates. It can be seen that EAPs based on the complete data exhibit the highest correlation, and Models 3 and 4 that treats missing data as incorrect show the lowest correlation. The squared correlation between the true person parameters and the person parameter estimates is a measure of reliability. A simulation study allows for comparisons of this *true* determination coefficient with

the model-based estimators of the reliability that are available in real applications. Using ConQuest, the EAP reliability as an overall estimator of the reliability, averaged over the distribution of the latent variable ξ, is available. This estimate of reliability is given by the ratio of the variance of the EAPs and the variances of the plausible values (Adams, 2005). The last column of Table 3 gives these reliability estimates based on the EAPs. Comparing the estimated reliability with the squared correlation of the true values of ξ and its estimators (column 4 of Table 3) reveals that Model 3 and to a lesser degree Model 4 overestimate the model-based reliability. For applications to real data this means treating missing data as answered wrong leads to biased and less accurate estimates and the estimated reliability will be spuriously high. All other models show only a slight overestimation with respect to the marginal reliability. Ignoring the missing data (Model 2) lowers the reliability, which is properly flagged by the lower estimated marginal reliability. Model-based approaches (see Models 5 to 7) show a slight improvement over Model 2 in terms of true and estimated reliability. Furthermore, the figure confirms the analytical derivations made above; the latent variable ξ is the same in Model 6 and 7.

Consequently, the EAPs are nearly identical in both models. Model 5 incorporates the latent regression $E\left[\xi \mid \bar{d}(U)\right]$, with the regression coefficient $\gamma_1 = \sigma\left[\xi, \bar{d}(U)\right] / \sigma^2(\xi)$. Given the measurement model of θ is true, with an increasing number of items, the covariance $\sigma\left[\xi, \bar{d}(U)\right]$ approaches the latent covariance $\sigma(\xi, \theta)$. Both Model 5 and Model 6 exploit the covariance of the response tendency and the latent ability. The only difference is that Model 5, the latent regression model, uses an observed statistic to quantify the response propensity. Therefore, it is expected that the EAPs are almost the same in both models, which can be verified by inspection of Figure 9.

It should be noted that the simple unidimensional IRT model (Model 2) that ignores the missing data is surprisingly robust in the case of the 30% missing data used in the simulation and a moderate correlation between ability and response propensity. To challenge the simple IRT model and to demonstrate the impact of systematic missing data on IRT model-based estimates, a second simulation was conducted with a stronger correlation between ξ and θ and with more missing data. The chosen conditions were $\rho(\theta,\xi) = 0.8$ and overall 50% missing responses in the data.

The presentation of results for this second simulation case study will be confined to the item parameter and person parameter estimates derived from IRT models. All CTT-based measures with respect to the items and persons exhibit the same shortcomings as already discussed above, but to a stronger degree.

The realized data in the second simulation exhibit an overall amount of missing data of 49.81%. It has to be noted that the omitted response rates differs across the items, depending on their difficulties. Table 4 gives the results on the accuracy of the estimated item parameters. The findings from the first simulation case are confirmed here with respect to Model 3. Treating missing data as incorrect responses leads to heavily biased item parameter estimates. Under both simulation conditions, Model 3 exhibits the largest expected bias. As can be seen in Figure 10, the item parameters are systematically underestimated.

Table 4

Bias, Standard Errors, and Mean Squared Error of the Item Difficulties of the Measurement Model of the Latent Proficiency Variable ξ Obtained in the Second Simulation With $\rho(\theta,\xi)$ = 0.8 and 49.81% Missing Data

Model	Item parameter estimates		
	Bias	*SE*	*MSE*
Model 1—complete data	-0.00519	0.00665	< 0.0001
Model 2—ignoring missing data	-0.13104	0.01672	0.01745
Model 3—missing = false	1.2026	0.1702	1.47521
Model 5—latent regression	0.01737	0.0159	0.00055
Model 6—between-item-multidimensionality	-0.01407	0.01546	0.00044
Model 7—within-item-multidimensionality	0.01081	0.01581	0.00037

The Models 5 to 7 take the missing data into account in an appropriate way and it appears that they are capable of reducing the bias considerably.

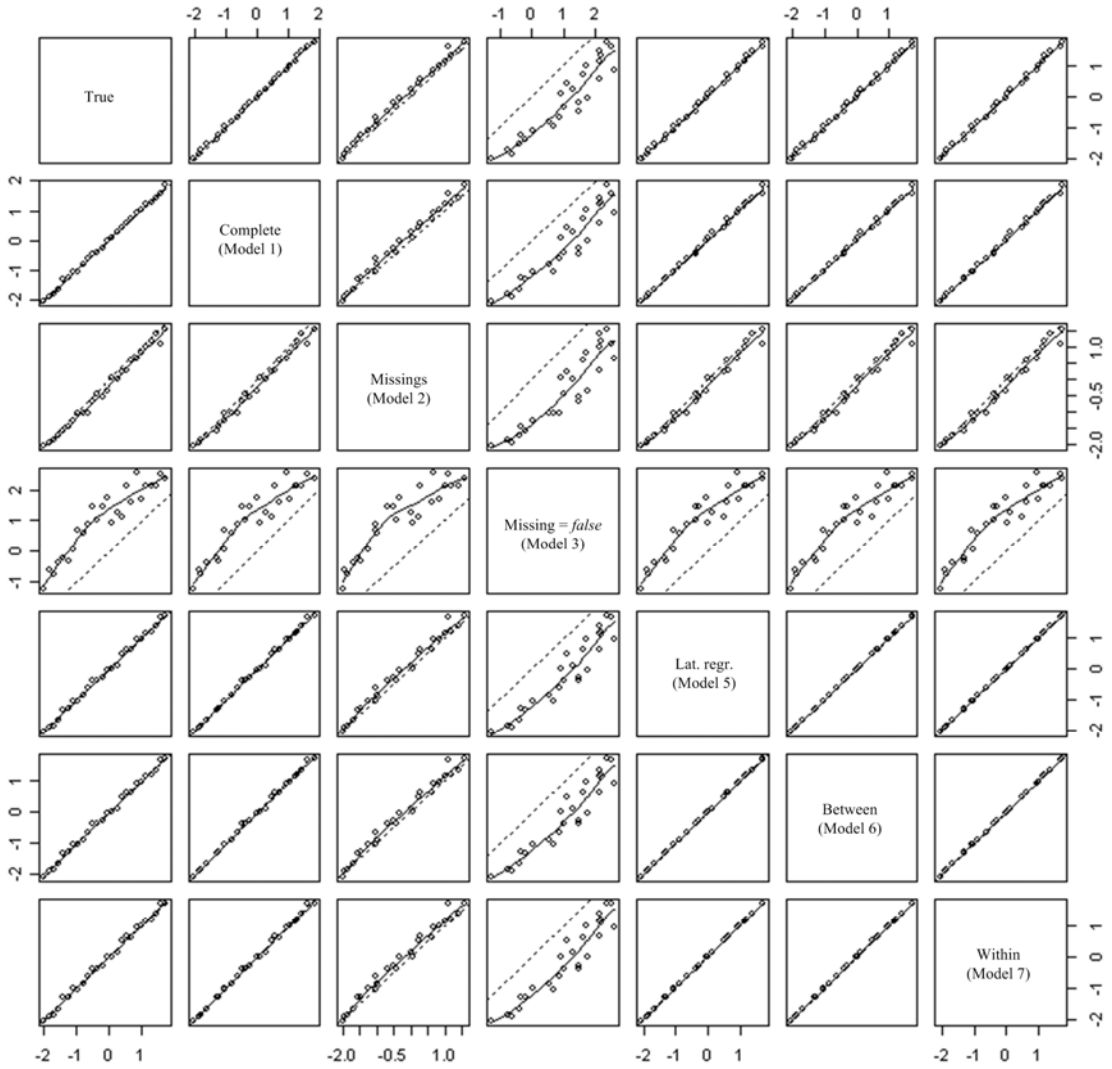

Figure 10. **Pair-wise comparisons of the estimated item difficulties of the second simulation with $\rho(\theta,\xi) = 0.8$ and 49.81% missing data. The first row/column illustrates the accuracy of the estimates of the different models by the comparison with the true items difficulties. The dotted line is the bisectric. The solid lines represent smoothing splines.**

Table 5 and Figure 11 summarize the results with respect to the EAP estimates of the different models in the second simulation case study. When ignoring omitted responses,

some simulated persons seem to benefit by skipping items that are too difficult for them. Due to this self-selection based on item difficulty, simulated respondents in the lower to middle range of ξ may achieve, on average, higher but potentially invalid estimates. Again, it can be argued that missing data and their treatment are a matter of test fairness.

Table 5

Bias, Standard Error, and Mean Squared Error of the EAP Estimators of the Latent Proficiency Variable ξ, Obtained in the Second Simulation With $\rho(\theta,\xi) = 0.8$ and 49.81% Missing Data

Model	Person parameter estimates (EAP)				
	Bias	*SE*	*MSE*	$r\left(\hat{\xi}_{EAP},\xi\right)^2$	$\mathrm{Rel}\left(\hat{\xi}_{EAP}\right)$
Model 1—complete data	-0.00759	0.43781	0.19174	0.812	0.812
Model 2—ignoring missings	-0.00938	0.62594	0.39189	0.615	0.636
Model 3—missing = false	-0.00847	0.53811	0.28963	0.733	0.828
Model 4—missing = false (2)	-1.42860	0.54460	2.33749	0.733	0.817
Model 5—latent regression	-0.00944	0.50222	0.23444	0.752	0.768
Model 6—between-item-multidimensionality	-0.04144	0.50395	0.25568	0.750	0.770
Model 7—within-item-multidimensionality	-0.01625	0.50302	0.25329	0.751	0.772

Note . $\mathrm{Rel}\left(\hat{\xi}_{EAP}\right)$ denotes the EAP-reliability defined as the ratio of the variance of the EAPs and the variance of the plausible values.

In general, the missing data mechanism is a source of variance in the EAP estimates. This is reflected by the lower correlation between the EAPs of Model 2 that ignores missing data and the true value of ξ compared to other approaches that incorporate response propensities, or by comparison with the case that is based on complete data. However, the simple unidimensional IRT Model 2 at least flags the lower correspondence between the ability variable and the EAPs by having a correspondingly low estimated reliability.

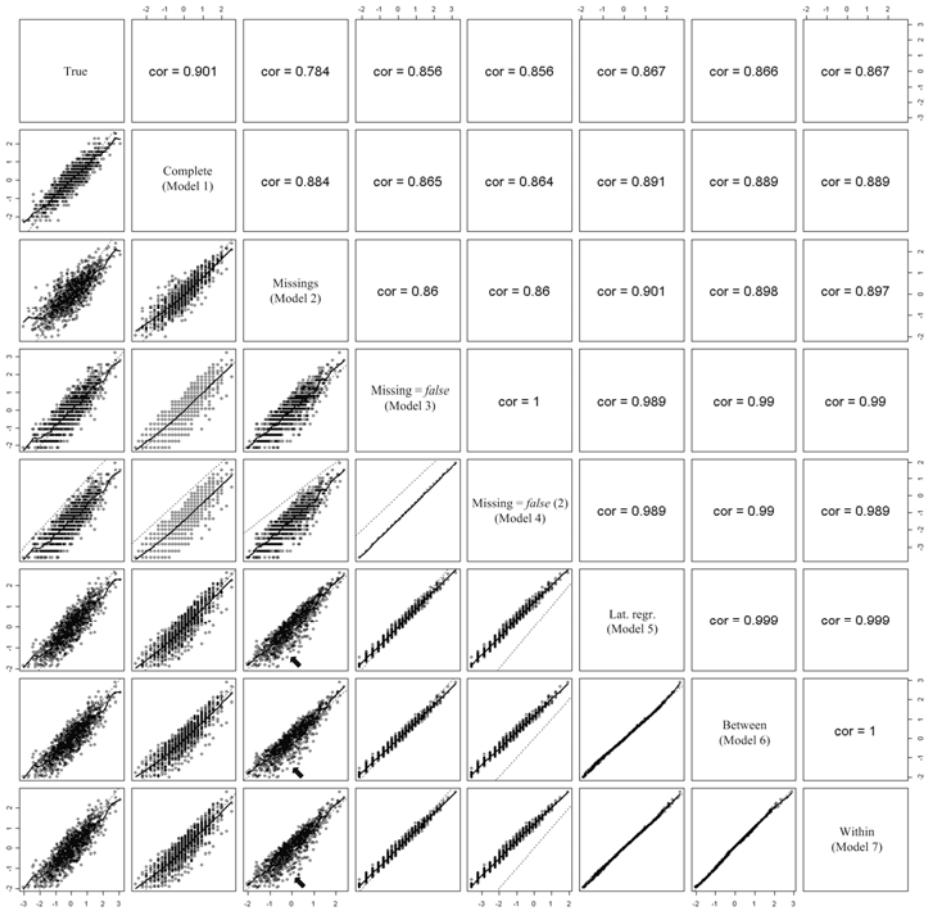

Figure 11. **EAP parameter estimates of ξ with $\rho(\theta,\xi) = 0.8$ and 49.81% missing data. True indicates the true person parameters ξ used to simulate the data in the second simulation with $\rho(\theta,\xi) = 0.8$ and 49.81% missing data. In the upper triangle the correlations are depicted. The dotted line is the bisectric and the solid line indicates a smoothing spline regression.**

Similar to the first simulation case, the reliability is noticeably overestimated when treating omitted responses as incorrect, as it is done in Models 3 and 4. The combination of handling missing data as "wrong" and using fixed item parameters from the model that ignores missing data leads to substantially underestimated person parameters. That can be contrasted with the overestimated marginal reliability. Compared with Model 2, it can be seen that Models 5 to 7 correct for the nonignorability of the missing data. On the one hand, that is reflected by the lower MSE of these models. On the other hand, it can directly be seen in the lower triangle of Figure 11. The arrows indicate the downward correction of

high person estimates in Model 3 by utilizing the missing data in Models 5 to 7. This leads to a reduction of error variance and increases the reliability of the EAPs.

The goal of the models compared above is their application to real data. We demonstrated the consequences of applying different models with simulated case-studies based on correlations of response propensities and abilities found in real data, as well as based on response rates as they relate to item difficulties found in real data. If the missing data mechanism used for the simulated data sets meets the underlying missing data mechanism of the omitted responses of the real PISA 2006 data, the missing data IRT models should show results for the PISA data similar to those observed in the simulation. IRT models are widely used in large-scale survey assessments of educational outcomes.

These low-stakes assessments typically suffer from a substantial amount of missing data, with the amount of omissions varying substantially from country to country, or between subpopulations within countries. Therefore the results presented here may be useful not only for the PISA assessment, but also for other large-scale surveys. However, we should also note that treatment of missing data differs across assessment programs, which limits the generalizability of the results. In PISA missing responses due to omissions are ignored in the item calibration but taken as answered not correctly in the following stages of operational analyses using the fixed item parameters from the previous calibration. In the next section, this handling of missing data and the more model-based treatments of missing data are compared using the PISA 2006 data.

Real Data Analysis: PISA 2006 Data

We used the PISA 2006 data of all OECD countries. The total sample size across these 30 countries is $N = 251{,}278$ cases. The PISA 2006 test consists of a total of 179 items; there are 48 items for mathematics, 28 for reading and 103 for science arranged into several booklets. Each booklet contains only a subset of the total item count, so that each student is measured with a relatively short test in the three domains: reading, math, and science. The observations were weighted with senate weights (i.e., following the operational procedures used in PISA, the sum of student weights was rescaled to 500 for each country for the analyses presented here.) It has to be noted that that in the original PISA data, different kinds of missing data were distinguished (Organisation for Economic Cooperation and

Development, 2009). In the analyses presented here, we collapse the missing data categories *not reached*, *omitted responses,* and *not codable* into one missing data category. This implies the assumption that these three categories of missing data are related to the latent proficiency variables of interest in similar ways. The responses missing by design, here missing due to the sparse booklet design, can be regarded as MCAR and were treated in that way. In the following sections, the term missing data denotes the missing values that are not caused by design factors but by examinee (non-)response behavior. The overall response rate is 90.21% across all OECD countries. This means that the total PISA data suffers from nearly 10% missing data due to some form of nonresponse. As Figure 12 reveals, the proportion of missing data varies substantially across the countries. Therefore, the missing data mechanism is at least related to covariates *Z*, in this context sometimes called *background variables*. Using the introduced terminology, it would initially seem that the missing data are at least MAR.

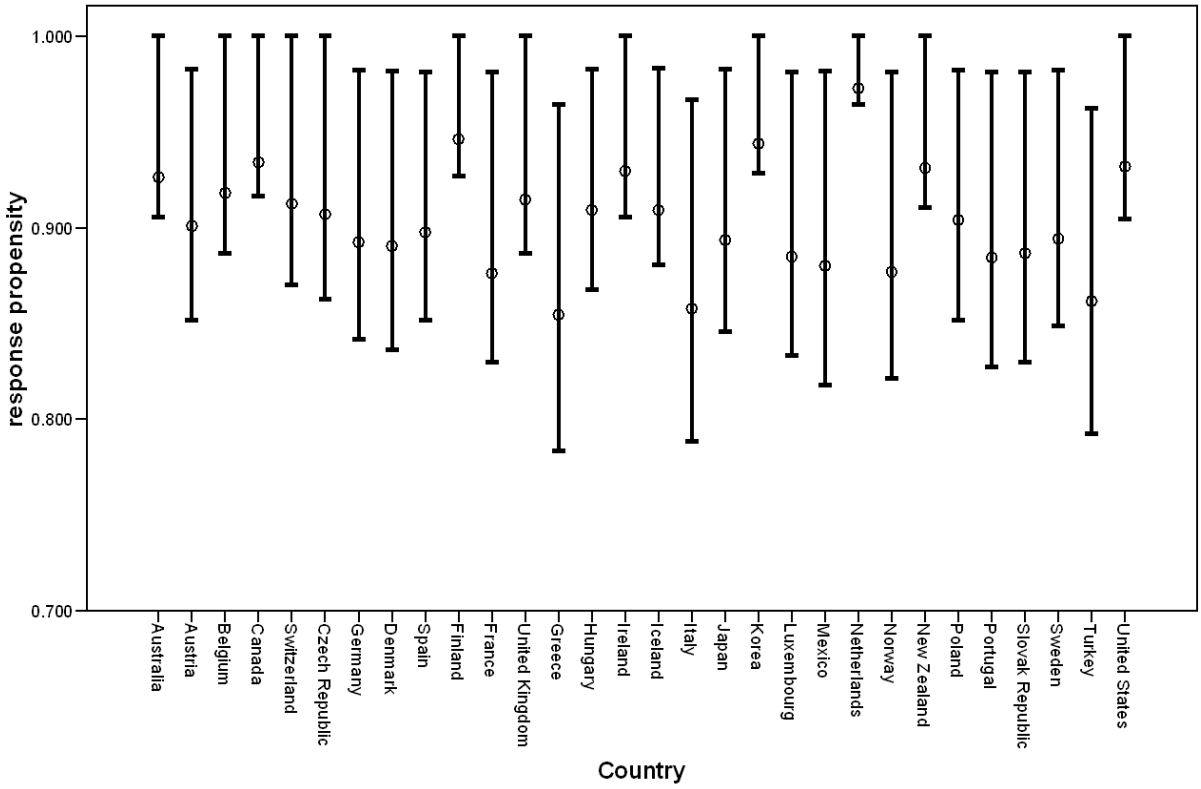

Figure 12. **Proportion of completed items in the PISA 2006 data for participating OECD countries. The circles indicate the means; the ends of the bars indicate the 25- and the 75-quantile.**

However, it is easy to find evidence that the missing data are also related to the proficiency of interest. There is a substantial correlation, $r = 0.398$, between the proportion correct $P^{+}(u)$ and the proportion of completed items $\bar{d}(u)$. Therefore, it is quite likely that the missing data mechanism is nonignorable. The proportion of completed items $\bar{d}(u)$ is the mean of the row vector in $\boldsymbol{D}$ that is associated with case u, and is the response rate of that person. It is a manifest estimator of the response propensity of a person. It was also found that the response rate $\bar{d}_i$ to an item Y_i depends on the standardized item mean $\bar{Y}_i^z$. This correlation is 0.33. Here, $\bar{d}_i$ is the column mean of the $\boldsymbol{D}$-matrix that corresponds to item Y_i. $\bar{Y}_i^z$ is computed by

$$\bar{Y}_i^z = \frac{1}{(C-1)\cdot n_{di=1}} \sum_{u=1}^{n_{di=1}} y_i(u),$$

and is the item mean computed with the sample size $n_{di=1}$ that responded to the item Y_i divided by the number of categories $C - 1$. In the case of dichotomous items, this is the proportion of correct answers in the observed responses to the item Y_i. As a result, the range of the standardized item means can only range between 0 and 1 even for partial credit items with $C > 2$ categories. Figure 13 (left) depicts the dependency between the standardized item means as a measure of the items difficulties and $\bar{d}_i$. Note that the proportion of completed items differs across the item types and is the highest for complex multiple-choice items and the lowest for open constructed-response (Figure 13, right).

The actual operational analyses of the PISA 2006 data were conducted with the Rasch model using *ConQuest*. Missing data were not treated with the model-based approaches as introduced earlier in this paper. The missing data were treated as wrong responses in the population modeling stages of the operational analyses. We reanalyzed the data using five different models, a subset of the models studied in the simulations previously presented. In order to avoid confusion, we use the same numbers to indicate the models as are used in the simulation. Obviously, Model 1 utilizing the complete data is not suitable, since the real data are incomplete. Furthermore, the within-item-multidimensional Model 7 was left out since Model 6 presents a more straightforward interpretation of the relation between response propensity and proficiency. This leaves Models 2 to 6, all of which were estimated for the PISA 2006 data. As outlined

above, the models are equivalent to the respective models in the simulation with the exception that they were adapted to the complexity of the real data. All real-data models include three latent ability variables: for mathematics, for reading, and for science. All models are multiple group models, with country as the grouping variable, or country by response propensity stratum in the case of Model 5.

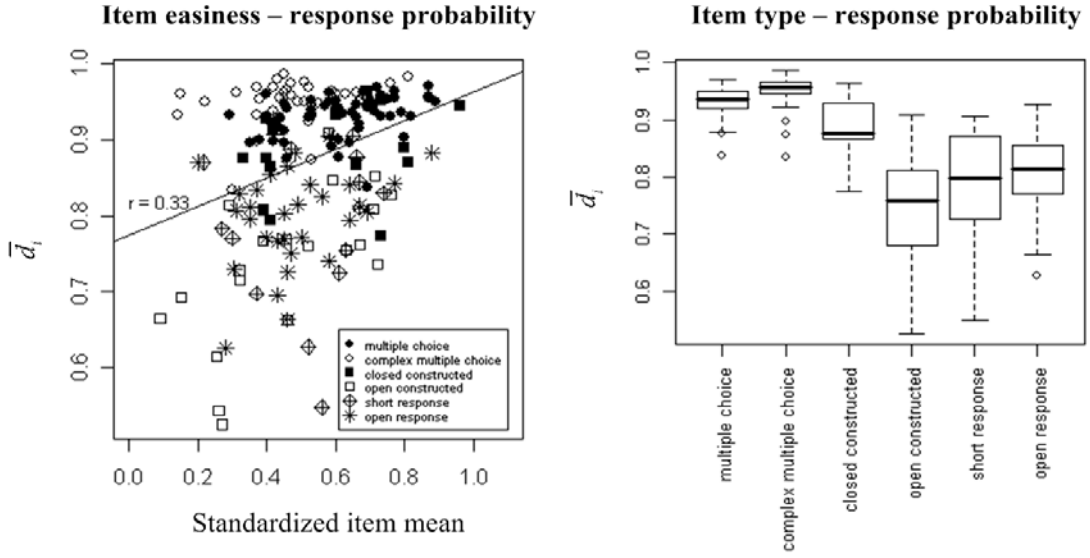

Figure 13. **Dependency between the standardized item means and the observed response rate of the items (left). The observed response rates depending on the item type (right).**

Since Models 5 and 6 both account properly for nonignorable missing data it is assumed that consistency between the estimated parameters of these models will be higher, compared to estimates obtained from Models 2 to 4.

Results

In this section, the item parameters will be considered first, followed by a comparison of the person parameter estimates. Due to the indeterminacy of the scale of (M-)IRT model parameters, some restrictions have to be introduced. The models are identified by fixing the sum of the item parameters per latent variable to one. To ensure the comparability of the estimates across the models, some standardization is needed that links the item and person parameter estimates across models. The method used here is based on the item difficulties. For each of the estimated models and each of the three proficiency dimensions $\xi_{k,}$ the item difficulties β_{ki} has been standardized in the following way

$$\beta_{ki}^{(z)} = \frac{\beta_{ki} - \hat{\mu}_{\beta_{k.}}}{\hat{\sigma}_{\beta_{k.}}} \ .$$

Where $\beta_{ki}^{(z)}$ is the standardized item difficulty of item Y_i. $\hat{\mu}_{\beta_{k.}}$ denotes the mean and $\hat{\sigma}_{\beta_{k.}}$ the standard deviation of all the item difficulties β_{ki} of the items Y_{ki} that indicate the latent variable ξ_k. The transformation of the slopes can be derived from the linear transformation of the item difficulties and is given by

$$\alpha_{ki}^{(z)} = \hat{\sigma}_{\beta_{k.}} \alpha_{ki}$$

Since, the item difficulties β_{ki} and the latent variables ξ_k share the same metric; the transformation of both is the same. Therefore the estimators of the latent variables ξ_k were linear transformed in the same way as the item difficulties

$$\xi_k^{(z)} = \frac{\xi_k - \hat{\mu}_{\beta_{k.}}}{\hat{\sigma}_{\beta_{k.}}} \ .$$

In this paper, only the conditional distributions of the person parameters within the countries are considered. So, the transformation has been conducted for the estimated means of each country g and each latent proficiency variable ξ_k. These group estimators will be compared across the models

$$\hat{E}\left(\xi_k^{(z)} \mid g\right) = \frac{\hat{E}\left(\xi_k \mid g\right) - \hat{\mu}_{\beta_{k.}}}{\hat{\sigma}_{\beta_{k.}}} \ .$$

Since we applied model variants based on a 2PL model, item difficulties and item discriminations can be compared across the five models. The results for comparing the item difficulties β_{ki} are summarized graphically in Figure 14. It can be seen that the item difficulties of Models 2 and 5 are nearly identical. The estimates of Model 6 are also very close to both of these models. Model 4 has been left out, because no item parameters were estimated in this model. Only the parameters of Model 3 that treated missing data as answered not correctly exhibit some deviations. If the item difficulties of Model 3 (the model that ignores the missing data mechanism) are subtracted from the item difficulties of Model 6 (between item-multidimensional model with latent response propensity), the resulting differences correlate with the response rates of the items by r = -0.570. Hence, 32.49% of the variance of the item difficulties across the

models can be explained by the proportion of missing data in the items ($F = 85.18$, $df_1 = 1$, $df_2 = 177$, $p < 0.001$). It is important to note that the simple Model 2 that ignores the missing data mechanism is very close to Models 5 and 6. This finding is consistent with the results from the simulation study discussed previously.

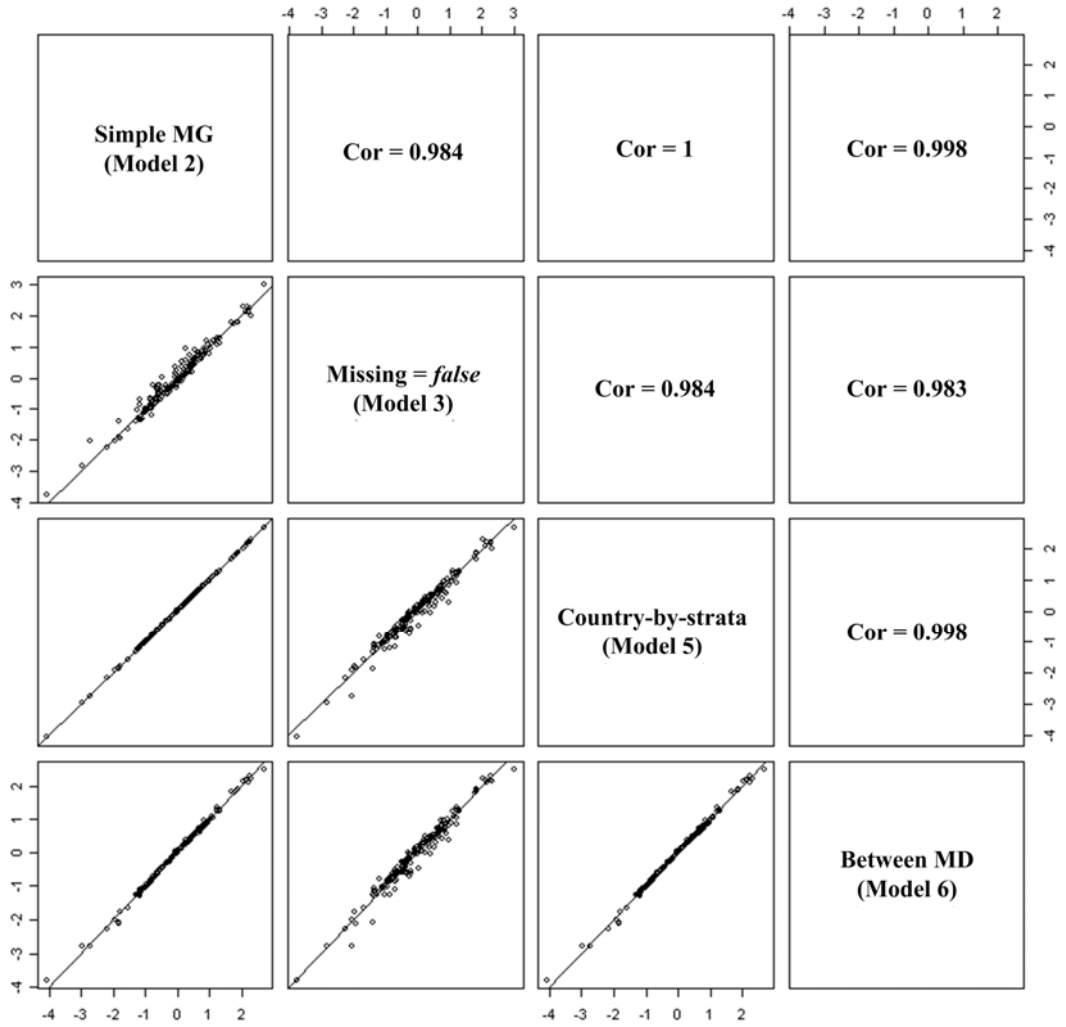

Figure 14. **Standardized item difficulties of the Models 2, 3, 5, and 6. Model 4 has been left out because the item parameters were fixed in that model. The solid line is the bisectric.**

Similar results can be found for the estimates of item discrimination parameters. The estimates are pretty much the same for Models 2 and 5, and the discriminations of Model 6 are very close to the discriminations from these two. Again the model that treats missing data

as wrong differs most from the other three. Figure 16 depicts these results. Compared to the item difficulties the differences between the estimated discrimination parameters of Models 2 and 5 seem to be less related to missing response rates. The determination coefficient is only 0.062 ($F = 11.7$, $df_1 = 1$, $df_2 = 177$, $p < 0.001$). The correlation is $r = -0.249$. In general, it can be stated that the variation of the estimated discrimination parameters across the models increases with the magnitude of the item discriminations.

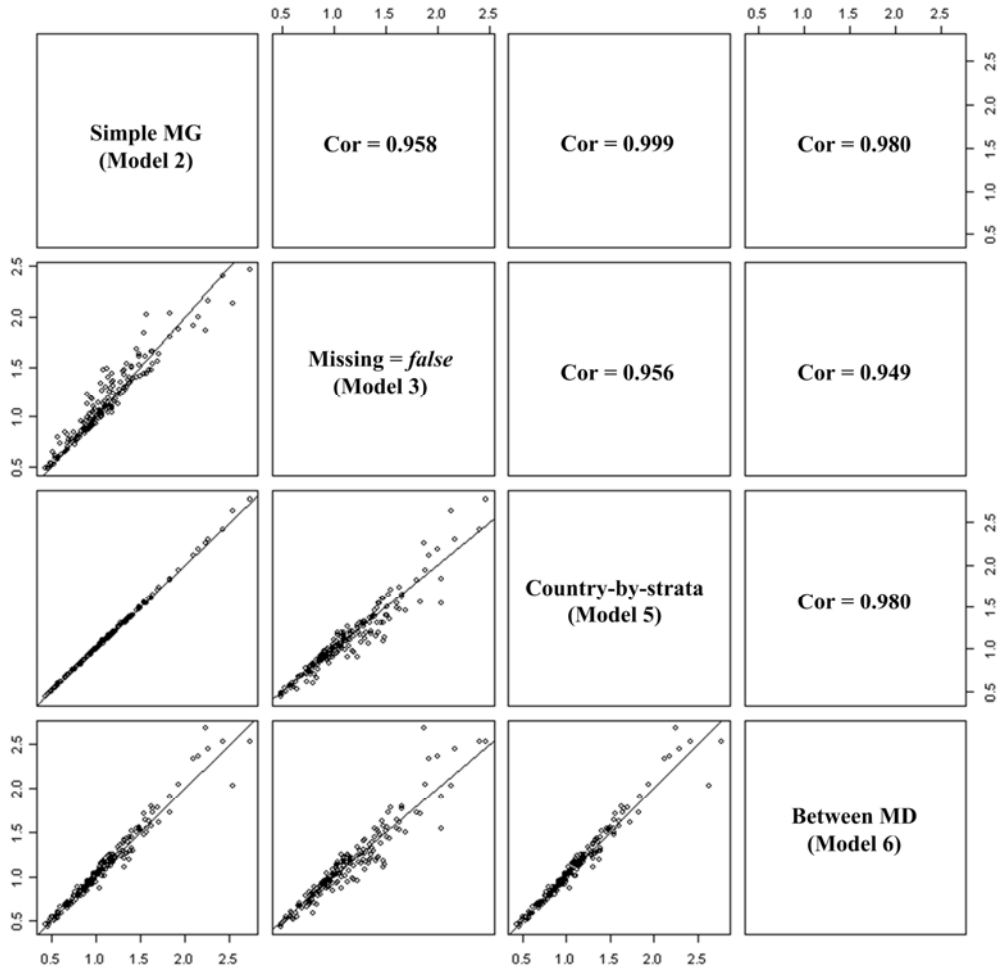

Figure 15. **Standardized item discrimination parameter estimates of Models 2, 3, 5, and 6: Model 4 has been left out because the item parameters were fixed in that model. The solid line is the bisectric.**

The person parameter estimates are now considered. As outlined above, the illustration of the results is confined to the estimated expectations $\hat{E}\left(\xi_k^{(z)} \mid g\right)$ of the three latent variables: mathematics, reading and science given the country variable indexed by *g*. Before the correspondence between the estimators across the models will be considered, the dependency between the response propensity and the latent proficiency variables needs to be discussed. As described above, a substantial correlation of $r = 0.389$ between the proportion of completed items and the proportion of correctly answered items has been found in the PISA 2006 data. It is expected that the correlation between the latent proficiency variable and the latent response propensity variable will be higher, since the observed correlation is attenuated due to unreliability. Table 6 shows these conditional correlations given the countries. The mean correlation between the latent abilities and the latent response propensities averaged across the countries is $\overline{r}_{math} = 0.433$ for mathematics, $\overline{r}_{read} = 0.434$ for reading and $\overline{r}_{sci} = 0.453$ for science. That means that 18.78%, 18.81%, and 20.49% of the variance of the latent response propensity variable can be explained by simple linear regression on the respective latent ability.

A careful look shows that the variance between the correlation coefficients is higher between the countries than between the scales within the countries. Thus, the country variable seems to be a moderator of the stochastic dependency between the response propensity and the latent ability variables.

Model 5 provides a different way of looking at the dependency between the latent proficiency variables and the latent response propensity variable. For each country there exists three groups. The first group is defined by a low average response rate and many missing responses, the second group has somewhat less missing data, and the third group has a very high response rate and therefore, very low proportion of not completed items. The positive correlation between the latent response propensity and the latent ability should be reflected in Model 4 by mean differences between the three groups within each country. Figure 16 shows the estimates of propensity group ability means by country and confirms this expectation.

Table 6

Estimated Conditional Correlation Between the Three Latent Ability Dimensions: Mathematics, Reading, and Science Given the Country and the Estimated Conditional Correlation Between the Three Latent Scales and the Latent Response Propensity, Given the Country

Country (g)	$Cor(\xi_k, \xi_t \mid g)$			$Cor(\xi_k, \theta_B \mid g)$		
	Mathematics Reading	Mathematics Science	Reading Science	Mathematics Response propensity	Reading Response propensity	Science Response propensity
AUS	0.800	0.874	0.840	0.468	0.476	0.509
AUT	0.802	0.923	0.847	0.508	0.516	0.524
BEL	0.763	0.830	0.859	0.498	0.506	0.495
CAN	0.872	0.823	0.866	0.404	0.424	0.429
CHE	0.830	0.834	0.891	0.483	0.497	0.496
CZE	0.797	0.910	0.841	0.473	0.459	0.479
DEU	0.841	0.865	0.918	0.534	0.555	0.535
DNK	0.736	0.843	0.852	0.435	0.472	0.469
ESP	0.864	0.863	0.922	0.413	0.448	0.433
FIN	0.740	0.832	0.876	0.320	0.392	0.353
FRA	0.678	0.881	0.878	0.470	0.456	0.492
GBR	0.848	0.841	0.864	0.489	0.521	0.530
GRC	0.729	0.767	0.918	0.356	0.374	0.386
HUN	0.811	0.916	0.822	0.460	0.459	0.449
IRL	0.774	0.876	0.833	0.485	0.498	0.532
ISL	0.666	0.863	0.887	0.394	0.420	0.406
ITA	0.625	0.853	0.865	0.372	0.364	0.398
JPN	0.847	0.858	0.864	0.470	0.457	0.483
KOR	0.856	0.813	0.803	0.423	0.424	0.445
LUX	0.734	0.839	0.825	0.446	0.440	0.473
MEX	0.706	0.805	0.784	0.289	0.267	0.281
NLD	0.706	0.868	0.913	0.405	0.424	0.412
NOR	0.884	0.834	0.857	0.431	0.437	0.461
NZL	0.873	0.860	0.891	0.462	0.517	0.523
POL	0.722	0.829	0.800	0.434	0.454	0.457
PRT	0.818	0.925	0.824	0.388	0.393	0.390
SVK	0.809	0.842	0.890	0.444	0.499	0.452
SWE	0.736	0.893	0.807	0.429	0.390	0.454
TUR	0.757	0.830	0.759	0.395	0.390	0.396
USA	0.077	0.842	0.081	0.422	0.082	0.437

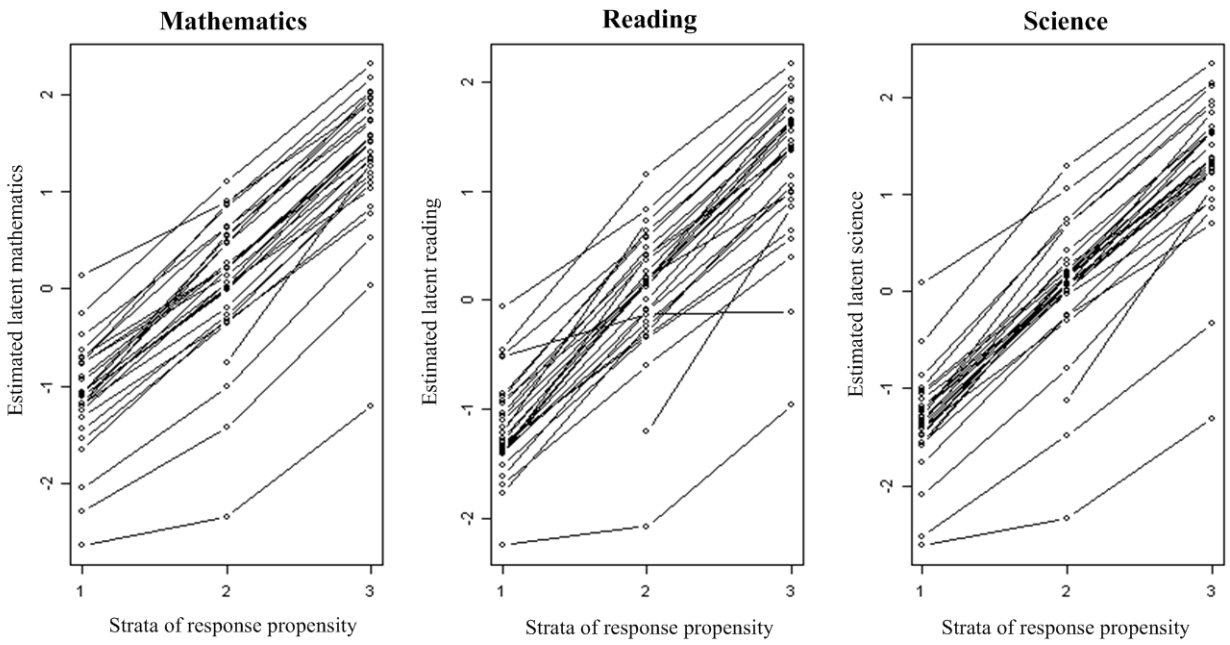

Figure 16. **Estimated means of the three different strata within the 30 OECD countries: Each three points that are connected by lines indicates a country.**

In order to demonstrate the effect sizes of the dependency between latent ability and the response propensity, the differences between the three groups for each country have been standardized. First, the three estimated means for each country were centered with respect to the mean of the second response propensity strata of that particular country. Afterwards the means were divided by the standard deviation of the third strata. So, in Figure 16, the country-specific mean differences between the adjacent strata are depicted in terms of standard deviations of the latent proficiency of the third strata. It can be seen, that the effect sizes reach up to absolute values of about 1 and slightly higher, between adjacent strata. even higher effect sizes result from the comparison of the first and the third strata within countries. Just as is the case with the conditional correlations $Cor(\xi_k, \theta_B \,|g)$ in Model 6, this means that the response rate and the average ability are correlated substantially. In other words, it appears that the missing data are not at random.

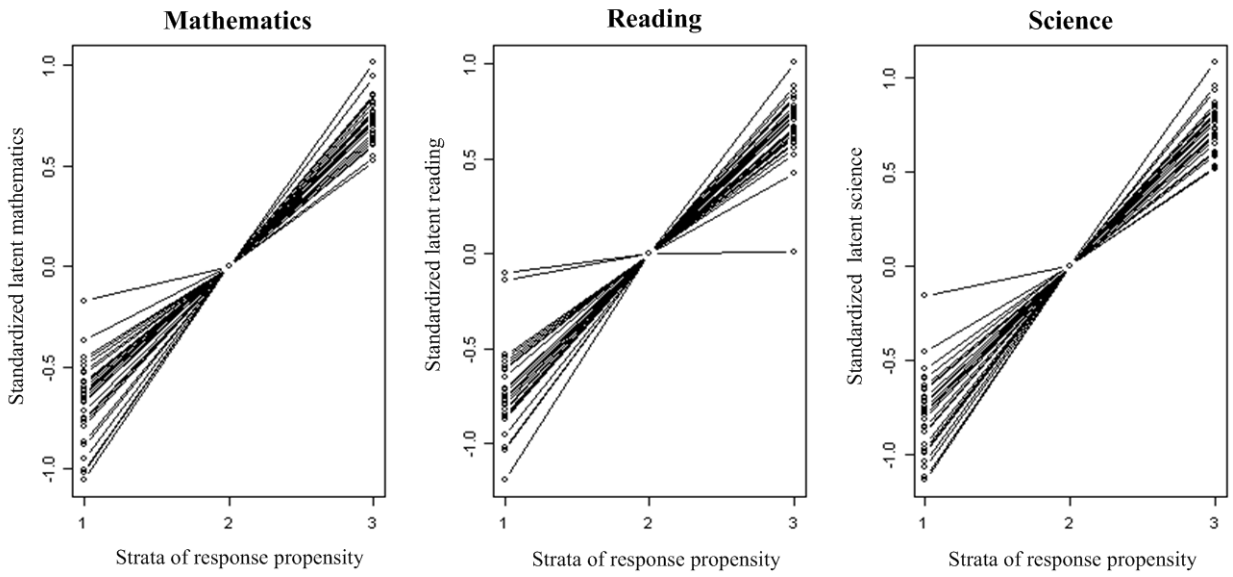

***Figure 17*. Effect sizes of the difference of the estimated means. All means are centered around the second strata within each country and divided by the standard deviation of the third strata. Hence the units of the *y*-axis are standard deviations of the third strata within the countries.**

Finally, the estimated conditional expectations $\hat{E}\left(\xi_k^{(z)} \mid g\right)$ of the three latent ability variables were compared across countries. It was expected that treating the missing data by recoding it into answered not correctly will lead to an underestimation of the person parameter estimates. This effect should be visible in the group means $\hat{E}\left(\xi_k^{(z)} \mid g\right)$. Figure 18 displays the results graphically. The different symbols indicate the three different latent variables. Circles indicate mathematics, squares mark reading, and triangles are indicative for science. Similar to the results of the item parameters, Models 2, 5, and 6 exhibit a high level of agreement. As hypothesized, Models 3 and 4 that regard missing data as answered not correctly show a systematic underestimation of the conditional means. The effect is even slightly higher for Model 4. Again, there is evidence that in the presence of nonignorable missing data, the simpler IRT Model 1 that ignores the missing data seems robust. The difference between the conditional means of Models 3 and 5 correlates with the average response rates of the countries by $r = 0.673$. Hence, 45.24% of the variance in the differences of the country means with respect to the latent proficiency variables can be explained by the proportion of missing data of the countries.

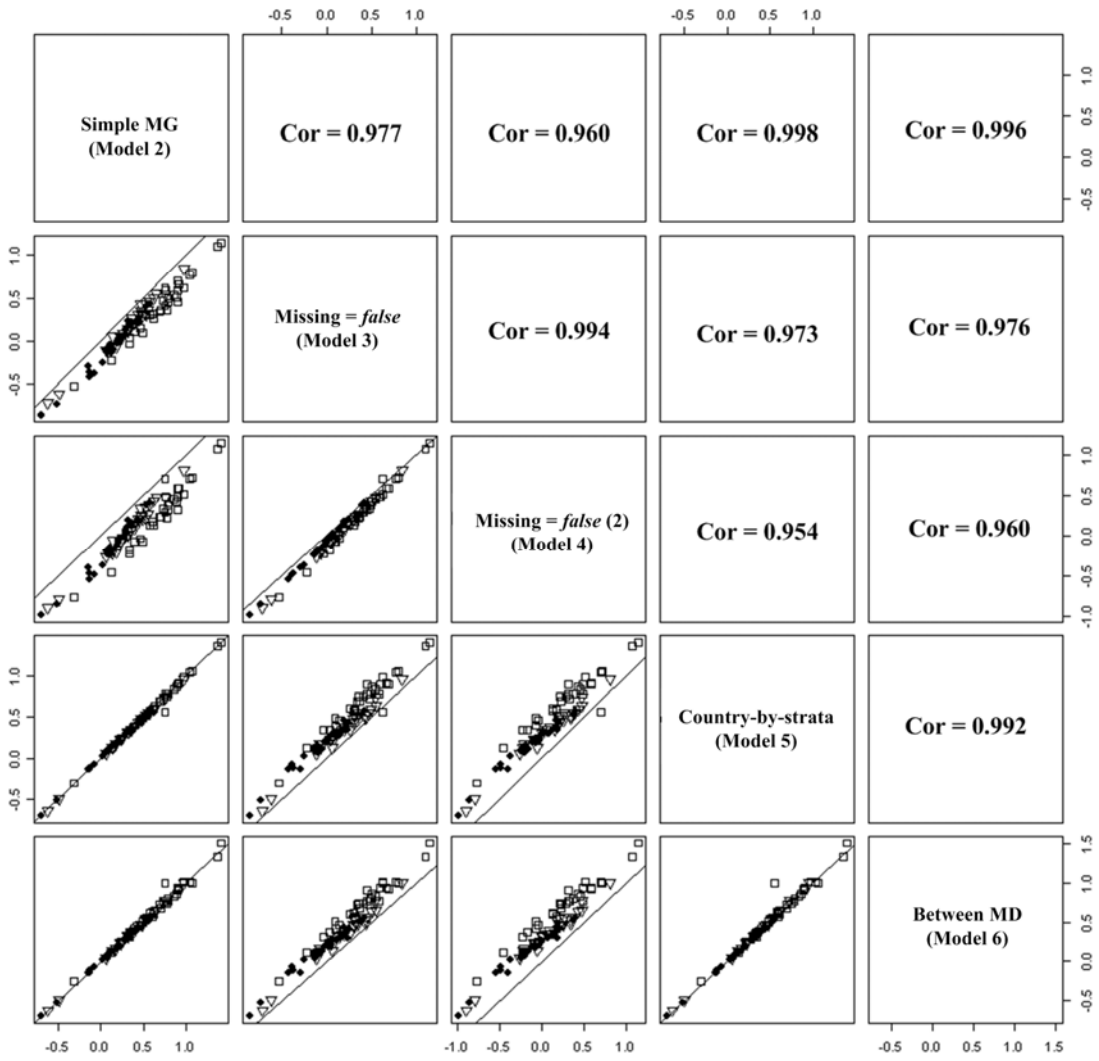

Figure 18. **Conditional means of the three latent variables mathematics, reading, and science across the five different models and given the countries. The different symbols marks the three variables (circles = mathematics, squares = reading, triangles = science). The red line is the bisectric.**

The presented results of the real data analyses so far provide no evidence that the Models 5 and 6 outperform the multidimensional multiple group Model 2. The results of the simulation studies showed that in cases of a substantial amount of missing data and a moderate correlation between the response propensity and the ability variable, the simple IRT model is quite robust. The bias is small even if missing responses are ignored. However, the case studies have shown that ignoring missing data lowers the reliability of the ability estimates. We expected to find improved

reliability estimates using the models that account for the missing data mechanism. Figure 20 (left) exhibits that the estimated reliabilities of Model 3 that treat the missing as answered wrong are reasonably higher than those of Model 2. Based on the simulation results, we consider the reliabilities of Model 2 to be overestimated and therefore to be biased. Also in accordance with the simulation results are the on-average higher reliability estimates from the between-item-multidimensional model (Figure 19, right). The previous theoretical considerations as well as the simulations support the consideration of these improved reliabilities as real. The use of the information of the missing data with respect to the latent ability improves the psychometric properties of the test.

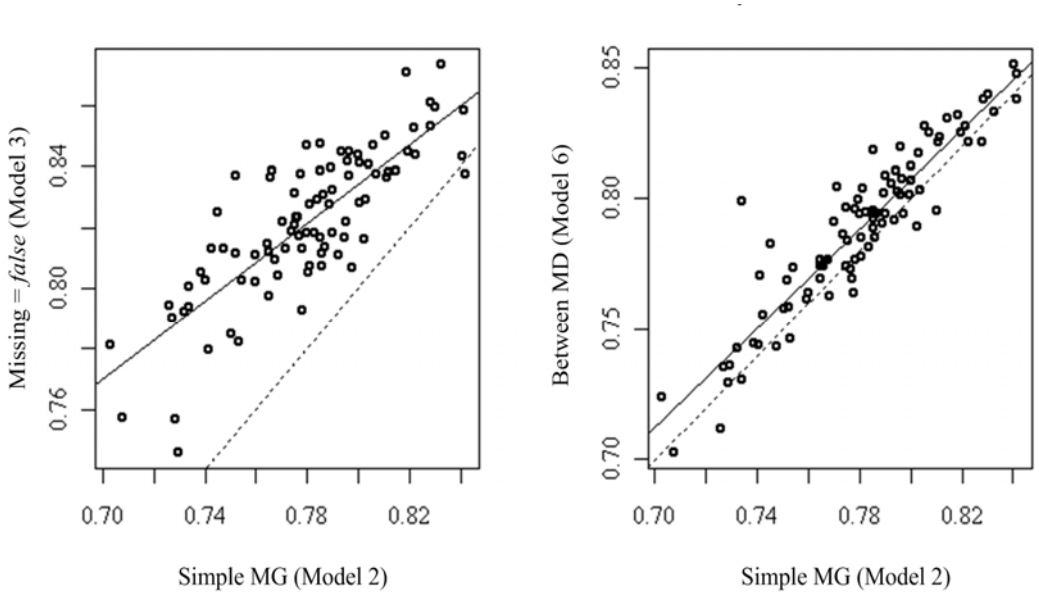

***Figure 19*. Comparison of the estimated reliabilities across the models given the countries and the scales: Comparisons between the reliabilities of the simple IRT model (Model 1) that ignores missing data with Model 2 that treats missing data as answered not correctly (left) and Model 6, the between item-multidimensional model, that accounts for the missing data mechanism (right). The solid line is a linear regression line. The dotted line is the bisectric.**

Discussion

In this paper, it has been shown that the presence of nonignorable missing data poses a threat to the reliability and the validity of test results reported using either CTT-based or IRT-

based methods. MNAR can lead to biased item and person parameter estimates. This has been demonstrated in a simulation case-study as well as with real data from PISA 2006. The effects of missing data that are not at random were demonstrated for the CTT-based measures such as the proportion correct (*P*+) of a person and the item means as a difficulty indicator of an item. The degree of bias depends on the strength of association between the latent proficiency and the response propensity and, obviously, the overall amount of missing data. The more missing-data that exists and the stronger its nonignorable nature is, the more biased the estimates will be. The results indicate that under a moderate proportion of missing data (30%), the person and item parameters estimates based on the IRT models are quite robust and give reasonably accurate estimates. But under high proportions of missing data (50%), person and item parameter estimates are seriously affected. The bias in CTT-based item and person parameters is already noticeable in the condition with 30% missing data. The item means $\overline{Y}_i(obs)$ are systematically overestimated, if the proportion of missing data is negatively related to the latent proficiency. It could be demonstrated that this overestimation increases with the difficulty of the items and results from the systematic dropout of responses in persons in the lower ability range. In other words, the item means are estimated on the basis of different subpopulations with respect to the latent variable that is the intention of measurement. The person parameters in CTT are the weighted or unweighted sum scores or functions of it. In the presence of missing data, the sum score is not appropriate, since the missing data were implicitly treated as answered not correctly. The mean of the completed items seems more appropriate and is often used in application. Unfortunately, the selection process that causes nonignorable missing data can lead to unfair comparisons between persons. It could be shown that persons tend to choose items that are more likely to be answered correctly with respect to their ability and to skip items with higher difficulties. That elevates the proportion of correct answers compared to a person with the same ability level but without omitted responses. In other words: Respondents administer a self-selected test that is closer to their proficiency and this item selection mechanism is, on average, beneficial. That the missing data are related to the latent ability is well known and might be the justification for tackling the problem by recoding the missing data into answered not correctly. However, considering the item mean and the proportion of correctly answered items in the case of dichotomous items, it was demonstrated that this procedure is also unfair. Treating missing data as wrong has an effect opposite to that from ignoring missing data, that is, it tends to

penalize respondents who actually might have solved the items. On average, those are the more proficient persons, who are less likely to gain from having missing data. Hence, this handling of the missing data does not correct satisfactorily for data MNAR. These findings underline that the occurrence of missing data as well as their treatment are a matter of test fairness.

Obviously, it would be best to find a procedure for test administration that avoids nonresponse completely. For large-scale survey assessments, this mainly means reducing the response burden, and increasing the motivation to respond in a manner such that respondents do their best in answering each item they are confronted with. In the absence of a method to achieve a perfect response rate, model-based approaches appear to provide ways to utilize the relationship between response propensity and proficiency in order to improve parameter and ability estimation. Using the latent variable models such as the ones specified in the IRT framework, the bias due to the nonignorable missing data can be reduced using models that take the missing data mechanism into account. Different models have been proposed in the literature. In the simulation case study, three model-based approaches were compared to the simple IRT model that ignores the missing data and the IRT model that treats the omissions always as wrong. In the between and the within MIRT models, a second latent variable is incorporated that captures the missing information, while the latent regression based missing data model uses a predictor based on the observed count of omitted responses to improve estimation of the proficiency variable. Using an IRT model that treats missing data as wrong confirmed that this treatment of missing data leads to heavily distorted item parameters and ungrounded overestimation of reliability. The results of the simulation case studies showed that these model-based approaches are equally suited to account for the nonignorable missing data. Note however, that the simple IRT models that ignores missing data shows relatively good performance under conditions of moderate amounts of missing data. The model-based approaches that incorporate a nonresponse variable of some kind outperform the simple IRT model clearly only under high rates of missingness. These models, however, have a secondary gain, in that they show the level of correlation between nonresponse and proficiency, and in that they improve the estimation of the reliability of ability estimates, and thus reduce uncertainty associated with estimators of proficiency distributions. It might be misleading to look only at correlations, overall biases, and means that summarize quantities. Statistical analysis strives for improvement of individual estimates as well as estimates of group-level distributions. For that reason, all relevant sources of information in the

data should be exploited in order to improve the diagnostic value of the estimates. As demonstrated, the model-based approaches of treating missing data adjusted the EAP ability estimates selectively, due to the pattern of missing data, and corrected for the unfair benefit of the systematically skipped items. So, even in the presence of only a few missing responses, there might be some test-takers with a high percentage of systematic missing data. For these particular cases, the EAP estimates based on models that adjust for the missing data mechanism are more trustworthy when compared to those based on models that ignore the missing data.

In the second part of our study we analyzed the data of PISA 2006. A total of 179 items measuring three latent variables were used in the assessment. The results of five models were compared for 30 OECD countries participating in PISA. Three of these models did not account for the stochastic but nonignorable nature of the missing data. As shown in the simulation study, it was confirmed that treating omitted responses as wrong produces item and person parameters that deviate from the other models. In the simulation study, we were able to confirm that treating missing data as wrong led to biases, while in the real-data analysis, we were only able to confirm that treating missing data as wrong leads to differences from the other approaches. First and foremost, the means of the person parameters and the item difficulties were affected. The bias was even larger when missing data were treated as wrong and additionally the item parameters were taken as fixed from an item calibration where the nonresponses were ignored.

As was found in the simulated case study, the simple model that ignores missing data produces results that are very close to the parameter estimates from the IRT models that incorporate the information on the missing data. Two such models were applied to the PISA 2006 data: The *between-item multiple group MIRT model* and a *multiple group model where the stratified response rates crossed with the country variable were used as a grouping variabl*e. The latter model emulates a latent regression model with the omission rate as a predictor for the latent ability variables. This approach is equally suited to account for systematic omitted responses as the between-item MIRT model that adds another dimension to the 3-dimensional model needed for PISA. Note that this yields a 4-dimensional IRT model with increased computational costs, while the latent regression missing-data model comes at virtually no cost. Operational analyses of large-scale surveys already involve high-dimensional latent regression models anyway, which would simply be augmented by the response propensity stratum as an additional predictor. The

effect sizes in Figures 16 and 17 show that the response propensity stratum is highly predictive of average ability in all 30 OECD countries in the PISA 2006 assessment.

We conclude that treating missing data as wrong appears to be the least desirable way to account for responses MNAR in large-scale surveys. Model-based approaches seem to provide a more appropriate way to account for nonignorable missing data. The estimation of model-based approaches allows the user to take the dependencies between response propensity and ability into account. The approaches that allow this to take place can be categorized as either *additional trait variable-* or *latent regression-based* approaches that promise to improve validity and comparability of large-scale assessments across participating countries in the presence of varying amounts of nonresponse.

References

ACER ConQuest. (2003). Generalised item response modelling software, Version 2.0 [computer software]. Hawthorn, Australia: ACER (Australian Council for Educational Research) Press: 2003.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162–172.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1):1–38.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

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

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (ETS Research Rep. No. RR-08-45). Princeton, NJ: ETS.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–162.

Holman, R., & Glas, C. A. W. (2005). Modelling nonignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*. *58*, 1–17.

Korobko, O. B., Glas, C. A. W., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, *45*(2), 139–157.

McLachlan, G. J., & Krishnan, T. (1996). *The EM algorithm and extensions* (2nd ed.). Hoboken, NJ: Wiley-Interscience

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of the Royal Statistical Society*, *Series A*, *163*(3), 445–459.

Organisation for Economic Cooperation and Development. (2009). *PISA 2006 technical report*. Paris, France: Author.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item nonresponse in attitude scales. *Journal of the Royal Statistical Society*, *Series A, 162*(2), 177–194.

Rijmen, F. (in press). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement.*

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

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

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

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Rep. No. RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2005). mdltm [computer software]. Princeton, NJ: ETS.

von Davier, M., DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In: J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 151–176). Cambridge, MA: Hogrefe & Huber.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam, the Netherlands: North Holland-Elsevier.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ACER ConQuest 2.0: General item response modelling software [computer software manual]. Camberwell, Australia: ACER Press.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Rep. No. RR-06-08). Princeton, NJ: ETS.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic models to NAEP data* (ETS Research Rep. No. RR-08-27). Princeton, NJ: ETS.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113–128.

Appendix A

A Note on Equivalent Models for Nonignorable Missing Data

The accuracy of the item parameters in the measurement model for θ will not be considered in this paper, since in real application the precise measure of the proficiency variable ξ and the item parameters of the associated measurement model are of primary interest. Some notes are provided here about the meaning of θ and the item parameter in the measurement model for θ. Figure A1 shows the relation between the item difficulties of Model 6 and 7 separately for the items Y_i and d_i.

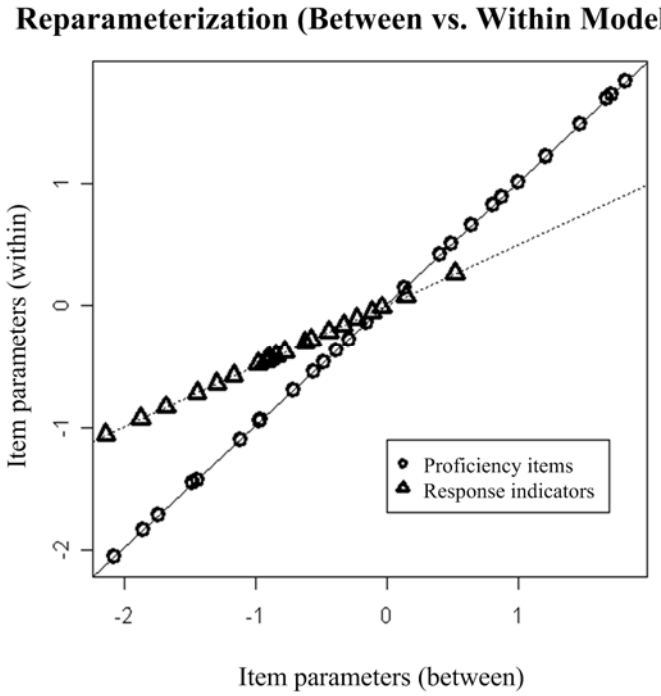

***Figure A1*. Relationship between the item parameters of the between-item-dimensional and the within-item-dimensional model.**

It can be seen that the item parameters are nearly identical for the items Y_i and the item parameters of d_i are linear functions of each other. That is quite different for the latent variable θ_B and θ_W in both models. It is important to note, that θ_B and θ_W have completely different meanings.

To demonstrate this difference, the logits of the manifest items in the model are used. In Model 6, the logits of the items Y_i and the latent response indicators d_i are

$$l(Y_i) = \xi - \beta_{Yi}$$
$$l(d_i) = \theta_B - \beta_{di}.$$

For Model 7 it follows that the logit of the items Y_i are the same but for the response indicators used in Model 7 are given by

$$l(d_i) = \xi + \theta_W - \beta_{di}.$$

It is easy to see that $\xi = l(Y_i) + \beta_{Yi}$. So, in both models, the latent proficiency variable is a translation of the item logits $l(Y_i)$. The latent variable θ_B is also a translation but of the logits $l(d_i)$ of the response indicator variables, and can be written as $\theta_B = l(d_i) + \beta_{di}$. Using these definitions, it can be shown that the latent variable θ_W can be rewritten as a translation of the difference of the logits of Y_i and the response indicator variables d_i

$$\begin{aligned}\theta_W &= l(d_i) - \xi + \beta_{di} \\ &= l(d_i) - \left(l(Y_i) + \beta_{Yi}\right) + \beta_{di} \\ &= l(d_i) - l(Y_i) + \left(\beta_{di} - \beta_{Yi}\right).\end{aligned}$$

The covariances and the correlations will differ across the models. The covariance $\sigma(\theta_B, \xi)$ in Model 6 is simply the covariance $\sigma[l(Y_i), l(d_i)]$ between the logits of Y_i and d_i. For Model 7 the covariance can be written as

$$\begin{aligned}\sigma(\theta_W, \xi) &= \sigma\left[l(d_i) - l(Y_i) - \left(\beta_{Yi} + \beta_{di}\right), l(Y_i) + \beta_{Yi}\right] \\ &= \sigma\left[l(d_i), l(Y_i)\right] - \sigma^2\left[l(Y_i)\right] \\ &= \sigma\left(\theta_B - \beta_{di}, \xi - \beta_{Yi}\right) - \sigma^2\left(\xi - \beta_{Yi}\right) \\ &= \sigma(\theta_B, \xi) - \sigma^2(\xi).\end{aligned}$$

So, this is a covariance between a difference variable and the minuend of that difference. This covariance is typically negative. In our simulation, the estimated correlation $r(\theta_B, \xi)$ in Model 6 is 0.551, whereas in Model 7 the correlation is $r(\theta_W, \xi) = -0.507$. Of course that does not mean that the tendency to omit responses is negatively associated with the latent ability

variable ξ. Furthermore, the variable θ_W has a different meaning and should not be called a latent response propensity.

Finally, the relation between the two variables θ_B and θ_W can be considered. Given, the regressions $E(\theta_B \mid \xi)$ and $E(\theta_W \mid \xi)$ are linear, it can easily be shown that θ_W is not simply a function of θ_B but it is linear regressive dependent on θ_B

$$\begin{aligned}\theta_W &= \alpha_0^{(W)} + \alpha_1^{(W)}\xi + \zeta_W \\ &= \alpha_0^{(W)} + \alpha_1^{(W)}\left(\alpha_0^{(B)} + \alpha_1^{(B)}\xi + \zeta_B\right) + \zeta_W \\ &= \underbrace{\alpha_0^{(W)}\alpha_0^{(B)}}_{\text{intercept}} + \alpha_1^{(W)}\alpha_1^{(B)}\xi + \underbrace{\left(\alpha_1^{(W)}\zeta_B + \zeta_W\right)}_{\text{residual}} .\end{aligned}$$

In the simulated data the estimated correlation $r(\theta_W, \theta_B)$ based on the expected a posteriori estimates is 0.386.

However, when using other models, other model specifications and identification rules are possible. In the Rasch model, all the discriminations are 1. Using the 2PL logistic model, the item discriminations between the measurement models for θ and ξ can differ. Insofar as no restrictions on the discrimination parameters are introduced, Model 7 is only identified if the covariance $\sigma(\theta_W, \xi)$ is fixed to be zero. In this case, θ_W is a residual of the regression $E(d_i \mid \xi)$. A detailed discussion of the consequences of different model specifications with respect to the meaning of the model variables and their relationships to other variables is far beyond the scope of this paper, but it is important to note that some authors denote the latent variable θ in different models equally as latent response propensity. This is misleading. Here we take the view that only the variable θ_B should be called a latent response propensity. Furthermore, in some papers it is not clearly pointed out that the model specification, as within or between item multidimensional models, does not depend on assumptions about the ignorability of the missing data. Therefore, in terms of model fit and the item and person parameters, the models are equivalent.

Appendix B

Proof

In this proof it is shown that $E(Y \mid d = 1) > E(Y)$ holds in general, given $\sigma(\theta,\xi) > 0$ and $P(d = 1) < 1$. That means the expectation of the variable Y under the occurrence of nonignorable missing data is always higher than under the absence of missing data, d is the associated dichotomous response indicator of Y; ξ is the latent ability variable measured by Y and θ the latent response propensity measured by d. The assumptions we make here are that the regressions $E(Y = 1 \mid \xi)$ and $P(d = 1 \mid \theta)$ are monotonically increasing functions of the respective predictor.

At first, it is shown that it is always true, that $\sigma(\theta,\xi) > 0 \Rightarrow \sigma(d,\xi) > 0$, under the assumptions made above, because $P(d = 1 \mid \theta)$ is a monotone function $f(\theta)$ of θ,

$$\begin{aligned}\sigma(d,\xi) &= \sigma\left[P(d=1\mid\xi)+\varepsilon,\xi\right] \\ &= \sigma\left[P(d=1\mid\xi),\xi\right] \quad .\end{aligned}$$

It follows directly, that $\sigma(\theta,\xi) > 0 \Rightarrow \sigma\left[P(d=1\mid\theta),\xi\right] > 0$ and $\sigma(\theta,\xi) > 0 \Rightarrow \sigma(d,\xi) > 0$.

The positive covariance implies that $E(\xi \mid d = 1) > E(\xi \mid d = 0)$. This is true because the difference can be expressed as a regression coefficient of the simple linear regression $E(\xi \mid d)$, with

$$E(\xi \mid d=1) - E(\xi \mid d=0) = \frac{\sigma(d,\xi)}{\sigma^2(d)}.$$

This expression will be always positive if $\sigma(d, \xi) > 0$. So, the conditional distributions $\xi \mid d = 1$ and $\xi \mid d = 0$ differ with respect to their central tendency. Given the derivations above, it follows that

$$\sigma(\theta,\xi) > 0 \Rightarrow E(\xi \mid d=1) - E(\xi \mid d=0) > 0.$$

The conditional distribution $\xi \mid d = 1$ of the latent ability given the item is answered is, on average, higher. The regression $E(Y \mid \xi)$ is a monotone function $f(\xi)$ of ξ. The expectation of the item $E(Y)$ is the expectation of the regression $E[E(Y \mid \xi)]$. For all monotone functions of ξ hold

$$E(\xi \mid d=1) > E(\xi \mid d=0) \Rightarrow E\left[f(\xi \mid d=1)\right] > E\left[f(\xi \mid d=0)\right].$$

It follows that

$$E\left[f(\xi \mid d=1)\right] > \left[f(\xi \mid d=0)\right] \Rightarrow E\left[E(Y \mid \xi, d=1)\right] > E\left[E(Y \mid \xi, d=0)\right]$$
$$\Rightarrow E(Y \mid d=1) > E(Y \mid d=0).$$

Now it is easy to show that $\mathrm{E}(Y \mid d = 1) \geq \mathrm{E}(Y)$. First, it holds that

$$E(Y) = E\left[E(Y \mid d)\right].$$

Because d is dichotomous, that means that

$$E\left[E(Y \mid d)\right] = E(Y \mid d=0)\left[1 - P(d=1)\right] + E(Y \mid d=1)P(d=1) \ .$$

The limits are given with

$$\lim_{P(d=1)\to 0} E(Y \mid d=0)\left[1 - P(d=1)\right] + E(Y \mid d=1)P(d=1) = E(Y \mid d=0)$$
$$\lim_{P(d=1)\to 0} E(Y) = E(Y \mid d=0)$$

$$\lim_{P(d=1)\to 1} E(Y \mid d=0)\left[1 - P(d=1)\right] + E(Y \mid d=1)P(d=1) = E(Y \mid d=1)$$
$$\lim_{P(d=1)\to 1} E(Y) = E(Y \mid d=1) \ .$$

So, $E(Y \mid d = 1) = \mathrm{E}(Y)$ if the probability of the response indicator $P(d = 1) = 1$. The probability $P(d = 1)$ approaches zero the more $E(Y)$ approaches the conditional expectation $E(Y \mid d = 1)$.
It follows that

$$E(Y \mid d=1) > E(Y \mid d=0) \Rightarrow E(Y \mid d=1) \geq E(Y) \ .$$

The proof underlines that the strength of the dependency between the ability and the response indicator expressed by the covariance $\sigma(d,\xi)$ and the overall response rate $P(d = 1)$ determines the difference between $E(Y)$ and $E(Y \mid d = 1)$. In application, this means that the mean

of the observed responses to an item Y is a systematically biased estimator in the presence of nonignorable missing data.