1

Introduction To Multilevel IRT Modeling

CONTENTS

1.1 Introduction

Most educational research data have a multilevel character. Typically, lower-level observations are nested in students at a second level, students nested in classes at a third level, classes nested in schools at a fourth level, and so forth. The multilevel structure implies multiple levels of analysis to account for differences between observations, between students, and between other higher-level units. A straightforward linear analysis is not possible due to dependencies between observations in each cluster, since the clustering leads to a violation of the common assumption of independently distributed observations.

When ignoring the nested structure of such data, aggregation bias (i.e. group-level inferences are incorrectly assumed to apply to all group members), also known as the ecological fallacy, is most likely to occur. Furthermore, the estimated measurement precision is most often biased. Partly in response to these technical problems, hierarchical linear or multilevel models emerged. They are characterized by the fact that observations within each group vary as functions of group-specific or lower-level (micro) parameters. These parameters may vary randomly across the population of groups as a function of second-level (macro) parameters. The multilevel model takes

the hierarchical structure into account, and variance components are modeled at each sampling level. As a result, homogeneity of results of students in the same class is accounted for since they share common experiences. Specifically, the multilevel model can describe relationships between one or more dependent variables, school and teacher characteristics (teacher's attitude, financial resources, class size), and student characteristics (achievements, social background).

After tackling computational problems, the appropriateness of multilevel models in educational research studies was shown by Aitkin and Longford (1986). From that time, multilevel modeling of hierarchically structured data received much attention and important contributions were made (e.g., Goldstein, 2003; Longford, 1993, Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). Besides technical innovations to estimate appropriate error structures, attention has also been focused on testing hypothesis of within-cluster, between-cluster, and cross-level effects.

1.1.1 Multilevel Modeling Perspective on IRT

With the increasing popularity in multilevel modeling, IRT models were synthesized with multilevel models in various ways. In the straightforward multilevel approach on item response theory modeling (Adams, Wilson, and Wu, 1997), the level of observations are defined as the first level, and the population distribution of students as the second level. Such an approach adheres to the multi-stage sampling design that is often used to collect educational data. When the data are collected through a complex multistage sampling design, standard analysis methods that rely on the assumption of independently and identically distributed observations are not suitable. Therefore, in the 1980s, techniques were developed for univariate multilevel response models (e.g., Bock, 1989; Raudenbush & Bryk, 1988). In the 1990s, relevant statistical techniques were developed, which supported the extensions of techniques for multilevel regression modeling of multivariate responses.

Others (e.g., Raudenbush & Sampson, 1999, Muthén, 1991), took a slightly different perspective and integrated the latent variable measurement model in a more general multilevel model. The general idea is that a multilevel design can include various latent variables at different levels. And when observing item responses as level-1 units, the response model automatically defines the lowest level of the multilevel model. Note that in educational research, latent variables are often measured at different levels. For example, for the purpose of accountability, latent variables are to be measured at the teacher, school, and district level, besides the student level.

IRT models were also viewed from the perspective of generalized linear mixed models (GLMM; see Skrondal & Rabe-Hesketh, vol. 1, chap. 30; Muthén and Asparouhov, vol. 1, chap. 31; De Boeck & Wilson, vol. 1, chap. 34). In specific, in the multilevel formulation of the Rasch model, the Rasch model is considered to define the lowest level of the GLMM. The presentation of IRT models as GLMMs has received much attention (e.g, Adams & Wilson, 1996; Adams, Wilson, & Wang, 1997; Kamata, 2001, 2007; Pastor, 2003; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; De Boeck & Wilson, 2004; Tuerlinkx & Wang, 2004). Various statistical computer

programs supports the estimation of GLMMs, which made the modeling framework directly accessible for different applications.

1.1.2 Bayesian Multilevel IRT Modeling

Developments in latent variable response modeling make it possible to model in a flexible way the response observations, which often contain several complex characteristics. First, response data are often sparse at the individual level. This sparsity complicates an estimation procedure for obtaining reliable estimates of individual effects. The individual response data are linked to many respondents and by borrowing strength from the other individuals' response data, improved estimates of individual effects can be obtained. In the same way, more accurate estimates can be obtained at an aggregate level using the within-individual data. Second, response data are often integer-valued. Responses are obtained as correct or incorrect, or obtained on a five- or seven-point scale. The lumpy nature of response data requires a special modeling approach since the standard distributional assumptions do not apply. Third, response data are often obtained in combination with other input variables. For example, response data are obtained from respondents together with school information, and the object is to make joint inferences about individual and school effects given an outcome variable. In a hierarchical modeling framework, different sources of information can be handled efficiently, accounting for their level of uncertainty.

The extension of an IRT model to more than two levels (i.e., item response defines level 1 and student level 2) is considered a multilevel IRT model. A typical example is educational survey data collected through multistage sampling, where the primary sampling units are schools, and students are sampled conditional on the school unit. Following Fox (2010), Fox and Glas (2001), and Aitkin and Aitkin (2011), among others, a complete hierarchical modeling framework can be defined by integrating the item response model with the survey population distribution. Besides item-specific differences, this multilevel item response model takes the survey design into account, the backgrounds of the respondents and clusters in which respondents are located. The heterogeneity between respondents is a typical source of variation that is modeled with the respondents' population distribution, which describes the between-individual and between-group variability.

The Bayesian modeling framework provides additional features (Congdon, 2001; Fox, 2010). First, it supports in a natural way extensions of common item response models, where prior models at separate levels are described to account for different sources of uncertainty, complex dependencies, and other sources of information. This flexibility is one of the strengths, which makes it possible to handle more complex sampling designs comprising complex dependency structures. Inferences can be directly made at different levels from posterior distributions of individual-level and higher-level parameters. Second, although the attractiveness of the Bayesian response modeling framework was recognized in the 1980s (e.g., Mislevy, 1986), Bayesian inference became feasible with the introduction of Bayesian computational methods such as computer simulation and Markov chain Monte Carlo (MCMC) techniques. The development of powerful computational simulation techniques induced a

tremendous positive change in the applicability of Bayesian methodology. The multilevel IRT model combined with powerful numerical simulation techniques make practical applications in educational test and survey research possible.

Multilevel IRT models accommodates the dependency typically found in hierarchical data, but also the estimation of latent variables and their relationships with explanatory variables at different levels. This chapter provides both a description and application of Multilevel IRT models. New developments and applications in this field will be demonstrated and directions for future research are given.

Besides the field of educational measurement, multilevel IRT models have been applied in other research fields. Van den Berg et al. (2007) showed how a multilevel IRT model can be applied in twin studies to account for measurement error variance that would otherwise be interpreted as environmental variance. They demonstrated that heritability estimates can be severely biased if analyses are simply based on sum scores. He et al. (2010) used the multilevel item response model to measure hospital quality and to assess its geographical variation. In this approach, hospital's quality was measured from different therapies, where a higher success rate corresponds to a better quality. The observations provided information about eligible patients receiving the therapy (coded one) or not receiving the therapy (coded zero). Patients were nested in hospitals, which were nested in geographical units. The IRT model enabled the measurement of a single quality score for each hospital, while accounting for differential measurement-specific weights. The multilevel component addressed the multilevel structure of the data.

1.2 Presentation of the Models

1.2.1 Multilevel IRT Model

Assume a multistage sampling design, where schools j $(j = 1,\ldots,J)$ are sampled, and subsequently students are sampled within each school j. Students' abilities are assessed using a test of I items. To ease the notation, a balanced test design is assumed such that each student p $(p = 1,\ldots,P)$ responds to each item i $(i = 1,\ldots,I)$.

Let U_{pji} denote the response of student p in school j to item i. For dichotomous items, a two-parameter IRT model describes the conditional probability of a correct response of student p to item i;

$$P\left\{U_{pji} = 1; \theta_{pj}, a_i, b_i\right\} = \Phi\left(a_i\left(\theta_{pj} - b_i\right)\right), \tag{1.1}$$

where $\Phi(.)$ represents the cumulative normal distribution function. Furthermore, the item discrimination parameter is denoted by a_i, the difficulty parameter by b_i, and the latent variable by θ_{pj}, which represents the student's ability.

The two-parameter IRT model in (1.1) defines the level 1 or observational level of the multilevel IRT model. The level-2 component describes the within-school distribution of abilities. Let level-2 explanatory variables be denoted by

$\mathbf{x}_{pj} = (x_{0pj}, x_{1pj}, \ldots, x_{Qpj})^t$, where x_{0pj} usually equals one. The level-2 model is represented by

$$\begin{aligned} \theta_{pj} &= \beta_{0j} + \ldots + \beta_{qj} x_{qpj} + \ldots + \beta_{Qj} x_{Qpj} + e_{pj}, \\ &= \sum_{q=0}^{Q} \beta_{qj} x_{qpj} + e_{pj} \end{aligned} \tag{1.2}$$

where the errors are independently and identically distributed with mean zero and variance σ^2_θ. The regression parameters are allowed to vary across schools. Level-3 explanatory variables are denoted by $\mathbf{w}_{qj} = (w_{0qj}, w_{1qj}, \ldots, w_{Sqj})^t$, where w_{0qj} typically equals one. The random regression coefficients defined in Equation (1.2) are considered to be outcomes in the linear regression at level 3,

$$\begin{aligned} \beta_{qj} &= \gamma_{q0} + \ldots + \gamma_{qs} w_{sqj} + \ldots + \gamma_{qs} w_{Sqj} + r_{qj}, \\ &= \sum_{s=0}^{S} \gamma_{qs} w_{sqj} + r_{qj} \end{aligned} \tag{1.3}$$

for $q = 0, \ldots, Q$. The level-2 error terms, $\mathbf{r}_j$, are multivariate normally distributed with mean zero and covariance matrix $\mathbf{T}$. The elements of $\mathbf{T}$ are denoted by $\tau^2_{qq'}$ for $q, q' = 0, \ldots, Q$.

Within each school j, the abilities are a linear function of the student characteristics $\mathbf{x}_j$ plus an error term $\mathbf{e}_j$. The level-2 random regression parameters $\boldsymbol{\beta}_j$ are assumed to vary across schools as a function of the school predictors $\mathbf{w}_j$ plus an error term $\mathbf{u_j}$. Within each school j, the matrix of explanatory data $\mathbf{x}_j$ is assumed to be of full rank. The level-2 and level-3 components can be represented by a single equation by filling in (1.3) into (1.2), and by stacking appropriately the matrices. Then, it resembles the general Bayesian linear model and allows $\mathbf{x}_j$ to be of less than full rank. Furthermore, not all level-2 parameters are necessarily random effects, where some of them can also be viewed as fixed effects (i.e., not varying across schools).

1.2.2 The GLMM Presentation

The Rasch model with components that describe the between student and between school variation is referred to as a multilevel Rasch model. The model permits the measurement of latent variables at different hierarchical levels, while accounting for the nested structure of the data at the level of students and higher levels. This model can be stated as a generalized linear mixed effects model (e.g., Adams et al., 1997; Kamata, 2001; Rijmen et al., 2003).

Consider the empty multilevel Rasch model, which does not have student or school predictors. Let π_{pji} denote the probability of endorsing item i for student p in school j. A logit or probit link function defines the relationship between the log-odds of the probability π_{pji} and a linear term with the item difficulty and ability parameter. Let indicator variable D_{pjk} equal one for student p when $k = i$, and zero otherwise.

Then, the level-1 model is represented by

$$\log\left(\frac{\pi_{pji}}{1-\pi_{pji}}\right) \quad = \quad \eta_{pj0} + \sum_{k=1}^{I} \eta_{pjk} D_{pjk},$$

where $\eta_{pjI} = 0$ to ensure that the design matrix is of full rank. The η_{pj0} can be interpreted as an overall effect across items and comprehends the ability level of of student p in school j. The level-2 and level-3 model is represented by,

$$\eta_{pj0} \quad = \quad \gamma_{00} + r_{0j} + e_{pj}$$

and $\eta_{pjk} = \gamma_{0k}$ for $k = 1, \ldots, I-1$. The error terms r_{0j} and e_{pj} are normally distributed and represent the between-student and between-school variation, respectively. The model can be extended with student and school predictors.

1.2.3 The Multiple Group IRT Model

Closely related to the multilevel IRT model is the multiple group IRT model. In some research studies, interest is focused in several specific groups. The selected groups are not considered to be sampled from a larger population. However, respondents are randomly sampled from each group. Bock and Zimowski (1997) proposed the multiple group IRT model, defining a group-specific population distribution, to handle the clustering of respondents in groups. This population distribution representing the clustered respondents completely specifies the distribution of respondents in each group. No assumptions are made about groups that are not selected. Inferences can be made with respect to the sampled groups but not to some higher level of population of groups.

When modeling the grouping structure of subjects using group-specific normal population distributions for the latent traits, the multiple group model can be seen as a natural extension of the two-parameter item response model. Azevedo, Andrade, and Fox (2012) generalized the multiple group IRT model. In a Bayesian modeling approach, other item response functions are considered such as the skew probit, logit, and the log-log. The multiple group latent variable distribution is characterized by a normal, Students t, skew normal, skew Students t, or finite mixture of normals.

This flexibility in item response functions and population distributions is parameterized by a mixture of different response functions ($l = 1, \ldots, L$) based on different cumulative distribution functions ($h = 1, ..., H$), and different latent trait distributions across items and groups. For a dichotomous response, the success probability of this generalized multiple group IRT model is given by

$$\begin{aligned} P\left\{U_{pji} = 1; \theta_{pj}, \boldsymbol{\xi}_i, \boldsymbol{\omega}\right\} \quad &= \quad \sum_{l=1}^{L} \prod_{h=1}^{H} F_{lh}\left(\eta_{pj}, \boldsymbol{\xi}_i, \boldsymbol{\omega}\right) \\ \theta_{pj} \mid \boldsymbol{\eta}_j &\sim G(\boldsymbol{\eta}) \end{aligned} \tag{1.4}$$

where cumulative distribution function F_{lh} has parameters, $\boldsymbol{\omega}$, and $G(\boldsymbol{\eta}_j)$ represents a

continuous population distribution function, where $\boldsymbol{\eta}_j$ denotes the population parameters of group j. The model includes the well-known one-, two- and three-parameter item response models using a probit, logit, loglog, Students t, skew probit, or skew probit link function.

1.2.4 The Mixture IRT Model

In the multiple group and multilevel IRT model, it is assumed that the response data are sampled from respondents nested in manifest groups. The groups are observed entities such as schools and countries. The nested structure leads to additional dependencies between response patterns. When the respondents are clustered but the clusters cannot be observed directly, a latent class model can be used. The latent class model can be used to capture the nesting of students in latent clusters and to identify the associated additional dependencies.

Following Rost (1997), when student p is classified to latent class g ($g = 1,\ldots,G$), the success probability of a correct response is given by

$$\begin{aligned} P\{U_{pgi}=1;\theta_{pg},\boldsymbol{\xi}_i,g\} &= \Phi\left(a_{ig}\left(\theta_{pg}-b_{ig}\right)\right) \\ \theta_{pg}\,|\,\mu_g,\sigma_g^2 &\sim N\left(\mu_g,\sigma_g^2\right), \end{aligned} \tag{1.5}$$

where θ_{pg} is the ability of student p in latent class g, a_{ig} and b_{ig} are the class-specific item discrimination and difficulty parameter, respectively. The class-specific mean and variance parameter are given by μ_g, σ_g^2, respectively.

The mixture IRT model has been used for detecting differential item functioning, differential use of response strategies, and effects of different test accommodations. When assuming measurement invariance, the mixture modeling approach is suitable to identify unobserved clusters of students. Vermunt (2003), and Cho and Cohen (2010), defined a multilevel latent class structure to model the latent clustering of schools and the clustering of students within schools. The mixture proportions at the school and student level can be modeled using explanatory information.

1.2.5 Multilevel IRT With Random Item Parameters

Item response data are cross-classified, which means that they are nested within students and nested within items. The IRT model uses item characteristic parameters to model the dependencies between observations due to the within-item clustering. Thus far, attention has been focused on the clustering of students. However, the item side of the multilevel IRT model needs to be correctly specified to make proper inferences.

The within-item correlation structure can be specified with an hierarchical prior. For the IRT model specified in Equation (1.1), a multivariate normal prior density for the item parameters can be specified as

$$(a_i,b_i)^t \sim N\left(\boldsymbol{\mu}_\xi,\boldsymbol{\Sigma}_\xi\right) I(a_i>0), \tag{1.6}$$

with hyper prior parameters

$$\begin{aligned} \boldsymbol{\Sigma}_{\xi} &\sim IW(\nu, \boldsymbol{\Sigma}_0), \\ \boldsymbol{\mu}_{\xi} \mid \boldsymbol{\Sigma}_{\xi} &\sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{\xi}/K_0). \end{aligned}$$

This prior assumes that the item characteristics are measurement invariant, meaning that they are equal across populations. As an extension to this multivariate normal prior, prior distributions for item parameters have been discussed that account for measurement variance. For a variant item, the item response function is not identical across populations.

Different applications have been discussed, where items are assumed to function differently across populations. Glas, van der Linden, and Geerlings (vol 1, chap. 26) discussed item cloning, where items are generated by a computer algorithm. Variation in item parameters have also been discussed by De Boeck and Wilson (2004; vol. 1, chap. 33) and De Jong et al. (2007). Verhagen and Fox (2013) discussed a longitudinal IRT model for ordinal response data, where items function differently over time, using random item prior distributions.

The prior in Equation (1.6) defines invariant item characteristics over groups. To allow for slight variations in item functioning over groups, random item parameters can be defined that allow for slight changes in item functions over populations. Let group-specific item parameters ($\tilde{a}_{ij}$ and $\tilde{b}_{ij}$) vary from the mean (measurement invariant) item parameters (a_i and b_i), and be distributed as,

$$\tilde{\boldsymbol{\xi}}_{ij} = \left(\tilde{a}_{ij}, \tilde{b}_{ij}\right)^t \sim N\left((a_i, b_i)^t, \boldsymbol{\Sigma}_{\xi}\right) \tag{1.7}$$

for $j = 1, \ldots, J$. Independent random item parameters can be defined when $\boldsymbol{\Sigma}_{\xi}$ is a diagonal matrix.

The random item parameter specification in the multilevel IRT model induces an identification problem. In each group, the group-specific mean ability and the mean test difficulty are not identified. Therefore, the mean test difficulty is restricted to be equal across groups such that between-group score differences are attributable to differences in ability and residual differences in item functioning. This identification rule has the objective to let the random item parameters explain between-group residual variance, such that the multilevel IRT model with measurement invariant item parameters is the theoretical optimal model. In the same way, item discriminations are allowed to vary across groups, where the average item discrimination is constant over groups. The random item discriminations explains residual between-group variation in item discrimination, where the preferred case is measurement invariant item discriminations but small fluctuations are allowed. Differences in item functioning across groups can be explained by background differences (e.g., culture or gender differences) as described in Verhagen and Fox (2013). It is also possible to allow for cross-classified differences in item characteristics when, for example, cross-national and cross-cultural response heterogeneity is present.

1.3 Parameter Estimation

The Bayesian modeling approach gives a natural way for taking into account all sources of uncertainty in the estimation of the parameters. The fully Bayesian framework results in a straightforward and easily implemented estimation procedure.

Therefore, the multilevel IRT model requires prior specifications of all model parameters. In Fox (2010), vague proper priors for the remaining model parameters are defined. That is, non-informative inverse gamma priors are specified for the variance components. An inverse Wishart prior is specified for the covariance matrix. Vague normal priors are specified for the remaining mean parameters. Then, a Gibbs sampling procedure can be used to estimate all model parameters. Following Albert (1992), an augmentation scheme is defined to sample latent continuous responses. The item parameters and multilevel model parameters can be sampled directly from the full conditionals given the augmented data, as described by Fox and Glas (2001), and Fox (2010). Furthermore, several packages in R and WinBUGS can also be used to estimate the model parameters. Cho & Cohen (2010) developed WinBUGS programs to estimate the multilevel IRT model, which includes mixture components.

In the GLMM presentation, various programs exist to estimate the model parameters. Tuerlinckx et al. (2004) compared the performance of different programs (GLIMMIX, HLM, MLwiN, MIXOR/MIXNO, NLMixed, and SPlus) and found overall similar results.

1.4 Model Fit

Posterior predictive checks provide a natural way to check assumptions of the model. Therefore, discrepancy measures need to be defined that provide information about a specific model assumptions. The extremeness of a fitted discrepancy measure given the data is evaluated using posterior predictive data, which are generated from their posterior predictive distribution. Discrepancy measures have been proposed to evaluate the assumption of local independence and unidimensionality. For an overview to posterior predictive model evaluation (see Sinharay vol. 2, chap. 19). Posterior predictive checks for evaluating IRT models have been proposed by Glas and Meijer (2003), Levy, Mislevy, and Sinharay (2009), and Sinharay, Johnson, and Stern (2006), among others.

Multilevel IRT models can be compared using the Deviance Information Criterion (DIC). The DIC is defined as

$$\begin{aligned} DIC &= D\left(\hat{\boldsymbol{\Omega}}\right) + 2p_D \\ &= -2\log p\left(\mathbf{y} \mid \hat{\boldsymbol{\Omega}}\right) + 2p_D \end{aligned}$$

where $\boldsymbol{\Omega}$ represents the multilevel IRT model parameters and $D\left(\hat{\boldsymbol{\Omega}}\right)$ the deviance evaluated at the posterior mean $\hat{\boldsymbol{\Omega}}$, and p_D represents the effective number of parameters and equals the posterior mean of the deviance minus the deviance evaluated at the posterior mean of the model parameters. When $\boldsymbol{\Omega} = \left(\boldsymbol{\xi}, \boldsymbol{\gamma}, \sigma^2_\theta, \mathbf{T}\right)$, the likelihood of interest of the multilevel IRT model can be presented as,

$$\begin{aligned} p\left\{\mathbf{u}; \boldsymbol{\xi}, \boldsymbol{\gamma}, \sigma^2_\theta, \mathbf{T}\right\} &= \prod_j \int_{\boldsymbol{\beta}_j} \left[\prod_{p|j} \int_{\theta_{pj}} \prod_i p\left(u_{pji} \mid \theta_{pj}, \boldsymbol{\xi}_i\right) \right. \\ &\quad \left. p\left(\theta_{pj} \mid \boldsymbol{\beta}_j, \sigma^2_\theta\right) d\theta_{pj} \right] p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right) d\boldsymbol{\beta}_j, \end{aligned} \tag{1.8}$$

such that the fit of random effects are not expressed in the likelihood.

1.5 Empirical Example

Data of 2003 from the Programme for International Student Assessment (PISA) of the Organisation for Economic Co-operation and Development (OECD) were analyzed to illustrate the multilevel IRT model. The PISA 2003 results and data can be found at *http://pisa2003.acer.edu.au*. Following Fox (2010), performances of Dutch students was investigated using various background variables. Furthermore, the random item parameter multilevel IRT model was used to investigate measurement invariance assumptions across Dutch schools.

1.5.1 Data

From the PISA 2003 study, the Dutch student results in mathematics were investigated using the multilevel IRT model. Student performance in mathematics was measured using 84 items. Students were given credit for each item they answered correctly. Although some items were scored with partial credit, for this analysis all item responses were coded as zero (incorrect) or one (correct). In PISA 2003 each student was given a test booklet with clusters of items, and each mathematics item appeared in the same number of test booklets. A number of 3,829 students across 150 Dutch schools were selected, where students with less than nine responses were not included in the present analysis.

1.5.2 Model Specification

To investigate individual and school differences in student performances, an empty multilevel IRT model was used. The following empty multilevel IRT model was used

to analyze the data,

$$\begin{aligned} P\{u_{pji}=1;\theta_{pj},\boldsymbol{\xi}_i\} &= \Phi(a_i(\theta_{pj}-b_i)) \\ \theta_{pj} &\sim N(\beta_{0j},\sigma^2_\theta) \\ \beta_{0j} &\sim N(\gamma_{00},\tau^2_{00}). \end{aligned}$$

The three levels of the model were specified to identify the with-student, between-student, and between-school variability. To account for measurement variance, item characteristics were considered to be random item parameters.

1.5.3 Results

The parameters of the multilevel IRT models were estimated using MCMC, as implemented in the package mlirt[1]. A total of 10,000 MCMC iterations were made, where the first 1,000 iterations were used as a burn-in. The multilevel IRT model was identified by fixing the mean and variance of the scale to zero and one, respectively.

In Table 1.1, the parameter estimates of the empty multilevel IRT model are given under the label Empty MLIRT. On the standardized ability scale, the between-student variance was .43 and the between-school variability around .61. The estimated intra-class correlation coefficient was around 59%. This represented the percentage of variability between math scores explained by differences between schools. In PISA 2003, the estimated intra-class correlation coefficient varied from country to country, with many countries scoring above the 50%.

The PISA 2003 results were computed using plausible values for the student's math abilities. Multilevel parameter estimates can be biased when point estimates are used as a dependent variable. The plausible values facilitate the computation of standard errors, while taking into account the uncertainty associated with the ability estimates. The plausible values were obtained as random draws from the posterior distribution of the ability parameters given the response data. Fox (2010, chap. 6) showed that the multilevel IRT parameters estimates and standard deviations were comparable to the multilevel model estimates using plausible values as outcomes.

The between-student and between-school differences in math performance were investigated using background variables. At the student level, female, place of birth (Netherlands or foreign), language (speaks foreign language most of the time), index of economic, social and cultural status, were used as explanatory variables. At the school level, the school-level mean index of economic, social and cultural status was used to explain variability between the conditional average school performances. In Table 1.1, the multilevel IRT model parameter estimates are given under the label MLIRT. It was concluded that male students performed slightly better than the female students. Native speakers performed better than non-native speakers with a migrant background. Students from more advantaged socio-economic backgrounds generally performed better. The school's index of economic, social, and cultural status had a significant positive effect on the school's average score.

[1] The Splus and R package mlirt are available at www.jean-paulfox.com

TABLE 1.1
Math performances of Dutch students in PISA 2003: Parameter estimates of the multilevel IRT models

	Empty MLIRT		MLIRT	
	Mean	HPD	Mean	HPD
Fixed part				
Intercept	-.04	[-.17,.09]	.02	[-.10,.14]
Student Variables				
Female			−.16	[-.22,-.11]
Foreign born			−.28	[-.38,-.17]
Foreign language			−.23	[-.34-,-.12]
Index			.15	[.12,.18]
School Variables				
Mean index			.39	[.08, .70]
Random part				
σ^2_θ	.43	[.40,.45]	.40	[.37,.42]
τ^2_{00}	.61	[.47,.76]	.49	[.38,.61]

FIGURE 1.1
Random item difficulty estimates of items one to five for the 150 Dutch schools in PISA 2003

The multilevel IRT analysis were performed given measurement invariant items. That is, the test was assumed to be invariant across schools. The multilevel IRT model with random item parameters was used to investigate whether small deviations in item functioning across schools would lead to a better model fit. Therefore, the empty multilevel IRT model was generalized with random item parameters, and school-specific item parameters were specified. Although the optimal preferred model was the measurement invariant multilevel IRT model, the more flexible multilevel IRT model with random item parameters could capture additional residual variance.

The estimated intra-class correlation coefficient was slightly lower and around 43%. However, the item parameter estimates hardly varied across schools and only small variations in item discriminations were detected. For the first five math items the average difficulty estimates were -.94, .17, .32, 2.21, -1.02, respectively. In Figure 1.1, the estimated item difficulties per item for each school were plotted for item one to five. It can be seen that there was almost no variation in item difficulty across schools. The school-specific difficulty estimates did not differ much and were not significantly different from the average (Dutch-specific) item difficulty.

1.6 Discussion

An overview is given of the multilevel IRT modeling framework, which generalizes the two-level IRT model with an additional level. The psychometric literature shows extensions where the student population distribution and/or the item population model is extended. In both cases, an extra clustering of students and/or items is modeled by an extra hierarchical level in the model. Generalizations have been made to model different types of response data and different types of clustering, among other things. The Bayesian modeling approach gives the possibility to use MCMC methods for parameter estimation. Those methods enable a a joint estimation procedure and via posterior predictive assessment a way to evaluate model assumptions.

The model has shown to be useful for school effectiveness research, where differences within and between schools are explored. Differences can be studied using the item response data as outcomes at level 1, and student abilities as outcomes at level 2, while accounting for the nested structure of the data. The typical nested structure in school effectiveness research can be easily modeled via multilevel modeling techniques.

Multilevel IRT models have been considered where the student variable of interest is unidimensional. When multiple student abilities are involved in producing the observed responses, a multidimensional IRT model can be specified such that multiple student abilities are considered to be outcomes at level 2. This requires a multivariate multilevel model at level 2. Such a modeling framework has been developed for modeling responses and response times to measure ability and speed of working (see van der Linden, vol. 1, chap. 29).

References

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice* (pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 1 47-76

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Aitkin, M. & Aitkin, I. (2011). Statistical modeling of the national assessment of educational progress. New York, Springer.

Aitkin, M., & Longford, N.T. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149,* pp. 1-43.

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

Azevedo, C.L.N., Andrade, D.F., & Fox, J.-P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational statistics and data analysis, 56,* 4399 - 4412.

Bock, R.D. (1989). *Multilevel analysis of educational data*. New York: Academic Press.

Bock, D.R., & Zimowski, M.F., 1997. Multiple group IRT. In van der Linden, W. J., Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York: Springer-Verlag.

Congdon, P. (2001). *Bayesian statistical modelling*. West Sussex: John Wiley & Sons.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35,* 336-370.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Jong, M.G., Steenkamp, J.B.E.M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34,* 260-278.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Fox, J.-P. & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27,* 217-233.

Goldstein, H. (2003). *Multilevel statistical models*. Third Edition. London, Edward Arnold.

He, Y., & Wolf , R.E., & Normand, S.-L. T. (2010). Assessing geographical variations in hospital processes of care using multilevel item response models. *Health Services and Outcomes Research Methodology*, 10, 111-133.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 7993.

Kamata, A., & Cheong, Y. F. (2007). Hierarchical Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models - Extensions and applications* (pp. 217-232). New York: Springer.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33,* 519-537.

Longford, N.T. (1993). *Random coefficient models*. Oxford, Clarendon Press.

Mislevy, R.J. (1986). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993-997.

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338-354.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16,* 223243

Raudenbush, S.W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd Ed.)*. Newbury Park, CA: Sage.

Raudenbush, S.W., & Bryk, A.S. (1988). Methodological advances in studying effects of schools and classrooms on student learning. *Review of research in education, 15,* 423-476.

Raudenbush, S.W., & Sampson, R.J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29,* 141.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185-205.

Rost, J. (1997). Logistic mixture models. In W.J. van der Linden & R. Hambleton. *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

Sinharay, S., Johnson, M.S., & Stern, H.S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298-321.

Snijders, T.A.B. & Bosker, R.J. (2011). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling (2nd)*. London: Sage.

Tuerlinckx, F., & Wang, W. C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75-109). New York: Springer.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate, W., Meulders, M. & De Boeck, P. (2004). Estimation and software. In P. De Boeck & M. Wilson (Eds.),*Explanatory item response models: A generalized linear and nonlinear approach* (pp. 343-373). New York: Springer.

Van den Berg, S.M., Glas, C.A.W., & Boomsma, D.I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics, 37,* 604-616.

Verhagen, J. & Fox, J.-P. (2013). Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine*. (doi: 10.1002/sim.5692).

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33,* 213-239.