1

Multilevel IRT Models with Covariates and Multiple Groups

CONTENTS

1.1 Introduction

Most educational research data have a multilevel structure. For example, the data collected may be nested in students as a second level, students in classes as a third level, classes in schools as a fourth level, and so forth. Such structures require multiple levels of analysis to account for differences between observations, students, and other higher-level units. Since the typical clustering of multilevel data leads to (marginally) dependent observations, separate linear analyses based on the assumption of independent identically distributed variables at each of these levels are inappropriate.

When the nested structure of such data is ignored, aggregation bias (i.e., group-level inferences incorrectly assumed to apply to all group members), also known as the ecological fallacy, may occur. Furthermore, the estimated measurement precision will become biased. Partly in response to these technical problems, hierarchical or multilevel modeling has emerged, which is characterized by the fact that observations

within each cluster are assumed to vary as a function of cluster-specific level parameters. In turn, these parameters may vary randomly across a population of clusters as a function of higher-level parameters. Multilevel modeling takes such hierarchical structures into account, adopting appropriate variance components at each level of sampling. As a result, homogeneity of the observations of, for instance, students in the same class due to their common experiences is accounted for. Also, multilevel models can be used to describe relationships between one or more dependent variables with teacher (e.g., attitudes), school (e.g., financial resources, class sizes), and student characteristics (e.g., achievements, social background).

Aitkin and Longford (1986) where the first to show the appropriateness of multilevel models for educational research and to tackle their computational problems. From then on, the idea of multilevel modeling of hierarchically structured data has received much attention, and important contributions have been made addressing such technical issues as the estimation of appropriate error structures and the testing of statistical hypotheses of within-cluster, between-cluster, and cross-level effects (e.g., Goldstein, 2003; Longford, 1993, Raudenbush & Bryk, 2002; Snijders & Bosker, 2011).

1.1.1 Multilevel Modeling Perspective on IRT

The increasing popularity of multilevel modeling has also affected item response theory (IRT) modeling in various ways. In the straightforward multilevel approach by Adams, Wilson, and Wu (1997; vol. 1, chap. 32), the responses to test items are accepted as first-level observations, students and items as second level, and, for instance, the population distribution of the students' parameters as the third level. Their approach reflects a multi-stage sampling design often used to collect educational data. As already noted, for data collected through such designs, a standard analysis relying on the assumptions of independently and identically distributed observations is inappropriate. In order to deal with this issue, univariate multilevel response models were developed (e.g., Bock, 1989; Raudenbush & Bryk, 1988), which later were extended to deal with multivariate response data.

A slightly different perspective was offered by the integration of latent variable measurement models into a more general multilevel model, for instance, by Muthén (1991) and Raudenbush and Sampson (1999). Their general idea was that a multilevel design can include various latent variables at different levels. Consequently, when item responses are observed, it is natural to adopt an IRT response model as first-level model in a more general multilevel framework.

IRT models have also been approached from the perspective of generalized linear mixed modeling (GLMM) (De Boeck & Wilson, vol. 1, chap. 34; Muthén & Asparouhov, vol. 1, chap. 31; Skrondal & Rabe-Hesketh, vol. 1, chap. 30). Separate attention has been given to the specific case of GLIMM formulations of extensions of the Rasch model (e.g, Adams & Wilson, 1996; Adams, Wilson, & Wang, 1997; De Boeck, & Kuppens, 2003; De Boeck & Wilson, 2004, vol. 1, chap. 33; Kamata, 2001, 2007; Pastor, 2003; Rijmen, Tuerlinckx, Tuerlinkx & Wang, 2004). Nowadays,

various computer programs support the statistical treatment of GLMMs, which has made the approach accessible for use in a large variety of applications.

Unlike regular latent variable modeling, modeling of response data leads to a few new requirements. First, response data are often sparse at the individual level, which makes it difficult to obtain reliable estimates of individual effects. However, the individual responses by one respondent are typically linked to many other respondents through the higher levels in the model, and by borrowing strength from them, improved estimates of individual effects can be obtained. Second, responses are often integer valued, for instance, obtained as correct-incorrect or on a five-point or seven-point scale. This lumpy nature of the responses prevents the use of standard statistical distributions and requires a special modeling approach. Third, response data may be obtained in combination with other input variables. An example is responses obtained from students together with school variables, where the objective is to make joint inferences about individual and school effects. Again, such cases with different sources of information require accounting for the uncertainty inherent in each of the sources and can be handled most efficiently in a multi-level framework.

1.2 Bayesian Multilevel IRT Modeling

Following Aitkin and Aitkin (2011), Fox (2010), and Fox and Glas (2001) among others, a complete hierarchical modeling framework can be defined by integrating item response models with survey models for population distributions. Besides item-specific differences, this hierarchical type of item response modeling can be used to account for the survey design, background of the respondents, and clusters in which respondents are located.

A Bayesian approach to multilevel IRT provides additional features. First, it supports a flexible way of incorporating prior knowledge to account for different sources of uncertainty, complex dependencies, and other sources of information at separate levels, with subsequent inferences possible at each of these levels different levels from the posterior distributions of their parameters. This option is one of the natural strengths of the approach, which makes it possible to handle sampling designs with complex dependency structures.

Second, although the attractiveness of the Bayesian response modeling framework was already recognized in the 1980s (e.g., Mislevy, 1986), it only became feasible with the introduction of such computational methods as computer simulation and Markov chain Monte Carlo (MCMC). The development of powerful computational simulation techniques has introduced a tremendous positive change in the applicability of Bayesian methodology. Combining multilevel IRT modeling with powerful computational simulation techniques makes practical application in educational test and survey research possible.

This chapter provides a description of multilevel IRT modeling and shows a few applications. Besides, a few developments are discussed and directions for future re-

search are given. Although our applications are from educational research, multilevel IRT models have been applied in other fields as well. For instance, van den Berg et al. (2007) showed how a multilevel IRT model can be applied to twin studies in genetics to account for measurement error variance that otherwise would have been interpreted as environmental variance. More specifically, they demonstrated that heritability estimates can be severely biased if the analyses are simply based on sum scores. He et al. (2010) used a multilevel response model to assess geographical variation in hospital quality. Their observations were of patients receiving or not receiving the therapy, where hospital quality was measured by the success rate of the therapy. Their patients were nested in hospitals which, in turn, were nested in geographical units. Their complete model included an IRT model that enabled them to measure the quality score of each individual hospital while accounting for differential measurement-specific weights. In addition, its higher levels addressed the hierarchical structure of their data.

1.3 Presentation of the Models

1.3.1 Multilevel IRT Model

Assume a multistage sampling design, for instance, where schools $j = 1,\ldots,J$ are sampled from a district and subsequently students are sampled within each school j. The students' abilities are assessed using a test of $i = 1,\ldots,I$ items. To simplify the notation, a balanced test design is assumed where each student $p = 1,\ldots,P$ responds to each item. Thus, let U_{pji} denote the response of student p in school j to item i.

For dichotomous items, the following two-parameter IRT model describes the 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, a_i denotes the item discrimination parameter, b_i its difficulty parameter, and the latent variable θ_{pj} represents the student's ability (Hambleton & van der Linden, vol. 1, chap. 2).

The two-parameter IRT model in (1.1) defines the first or observational level of modeling. The second level explains the within-school distribution of the abilities by variables denoted as $\mathbf{x}_{pj} = (x_{0pj}, x_{1pj}, \ldots, x_{Qpj})^t$, where x_{0pj} usually equals one. The level-2 model is

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

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. Similarly, level-3 explanatory variables can be adopted, which we denote as $\mathbf{w}_{qj} =$

$(w_{0qj}, w_{1qj}, \ldots, w_{Sqj})^t$, where w_{0qj} typically equals one. The random regression coefficients defined in (1.2) are now considered to be the result of 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 $q = 0, \ldots, Q$, with level-2 error terms, $\mathbf{r}_j$, assumed to be multivariate normally distributed with mean zero and covariance matrix $\mathbf{T}$. The elements of $\mathbf{T}$ are denoted as $\tau^2_{qq'}$ for $q, q' = 0, \ldots, Q$.

Thus, within each school j, the abilities are modeled as a linear function of the student characteristics $\mathbf{x}_j$ plus an error term $\mathbf{e}_j$, where the matrix with the explanatory data $\mathbf{x}_j$ is assumed to be of full rank. Further, 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 $\mathbf{u_j}$. Both these level-2 and level-3 equations can be reduced to a single equation by substituting (1.3) into (1.2), stacking their matrices appropriately. The result 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; some of them can also be viewed as fixed effects (i.e., not varying across schools).

1.3.2 GLMM Presentation

A multilevel Rasch model can be reformulated as a generalized linear mixed effects model (e.g., Adams et al., 1997; Kamata, 2001; Rijmen et al., 2003). Consider an unconditional multilevel Rasch model that does not have any student or school predictors. Let π_{pji} denote the probability of student p in school j answering correctly to item i. A logit link function can be used to describe the relationship between the log-odds of the probability π_{pji} and a linear term with the item difficulty and ability parameter. In a more general formulation, let $\eta_{pj0}, \ldots, \eta_{pjI}$ denote the parameters of this linear term. Then, the level-1 model is represented by

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

where indicator variable D_{pjk} for student p will equal minus one when $k = i$ and zero otherwise. A constraint $\eta_{pjI} = 0$ is given to ensure that the design matrix is of full rank.

The random intercept η_{pj0} can be interpreted as the ability level of student p in school j. The effects $\eta_{pj1}, \ldots, \eta_{pj(I-1)}$ are constrained to be fixed across students such that they represent item effects. The η_{pjk} represents the offset of the kth item from the Ith item.

The variation in ability across students and schools is represented by a level-2 and level-3 model, which can be summarized as

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

where γ_{00} represents the average level of ability in the population. The error terms r_{0j} and e_{pj} represent the between-student and between-school variation, respectively; they are assumed to be normally distributed. Again, if necessary, the model can be extended with student and school predictors.

1.3.3 Multiple Group IRT Model

The notion of a multilevel IRT model is closely related to that of a multiple group model. In some studies, the interest is in more than one specific group. Although the respondents are randomly sampled from these groups, the groups themselves are not considered to be sampled from any larger population. For this case, Bock and Zimowski (1997) proposed a multiple group IRT model that included group-specific population distributions to handle the clustering of respondents into groups. Their model allows inferences made with respect to each of the sampled groups but not to some higher-level population across all groups.

Azevedo, Andrade, and Fox (2012) generalized the multiple group IRT model following a Bayesian approach with item response functions allowed to be skew probit, logit, or log-log functions (Albert, vol, 2, chap. 1) and a multiple group latent variable distribution that can be represented by a normal, Student's t, skew normal, or skew Student's t distribution, or any finite mixture of normal distributions. This flexibility of choice of response functions and population distributions is obtained by parameterizing mixtures $l = 1,\ldots,L$ of different response functions based on different cumulative distribution functions $h = 1,...,H$, with possibly different latent trait distributions across items and groups. For dichotomous responses, the success probability of this generalized multiple group IRT model is

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

where cumulative distribution function F_{lh} has parameters $\boldsymbol{\omega}$ and $G(\boldsymbol{\eta}_j)$ represents a continuous population distribution function with parameters $\boldsymbol{\eta}_j$ for group j. The model encompasses the well-known one-, two- and three-parameter item response models, each with a choice from the earlier mentioned link functions.

1.3.4 Mixture IRT Model

Both multiple group and multilevel IRT model assume response data sampled from respondents nested in manifest groups. When the respondents are clustered but the clusters cannot be observed directly, a latent class model can be used. The approach then captures the nesting of students in latent clusters as well as the dependencies between the response vectors it has created.

Following Rost (1997), the success probability of a correct response can now be

modeled as

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

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

This mixture IRT model has been used for detecting differential item functioning, differential use of response strategies, and the effects of different test accommodations. Assuming measurement invariance, the mixture modeling approach is suitable to identify unobserved clusters of students. Cho and Cohen (2010) and Vermunt (2003) used a multilevel latent class model to identify latent classes of students that are homogeneous with respect to item response patterns, while accounting for the manifest clustering of students in schools.

1.3.5 Multilevel IRT With Random Item Parameters

Item responses are nested both within students and items. Thus far, our attention was exclusively on the clustering of students. However, the item side of the multilevel IRT model needs to be correctly specified as well to make proper inferences.

The item characteristics are specified to be normally distributed around the average item characteristics. Furthermore, the item parameters of each item are assumed to be correlated. For the IRT model specified in (1.1), a multivariate normal prior density for the item parameters $i=1,...,I$ 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 prior hyperparameters

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

The specification amounts to the assumption of all item response functions being invariant across populations of test takers. Alternatively, different assumptions have been made to allow for functions that do vary across such populations. Such kind of 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 prior distributions for the items. Glas, van der Linden, and Geerlings (vol 1, chap. 26) discussed a different application, where items are generated by a computer algorithm. They modeled variation in item characteristics of clone items in sets of similar items, which only differ by surface characteristics.

To allow for variation in item functioning over populations, random item parameters can be defined that allow for changes in the response functions of the items over

them. Let population-specific item parameters $\tilde{a}_{ij}$ and $\tilde{b}_{ij}$ be distributed as,

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

where a_i and b_i are means across the populations $j = 1, \ldots, J$, and independent random item parameters are defined by choosing $\boldsymbol{\Sigma}_{\tilde{\xi}}$ to be diagonal. However, this random item parameter specification introduces an identification problem. For each population, both its mean ability and the mean item difficulty is not identified. Therefore, we may decide to restrict the mean item difficulty to be equal across populations, so that possible score differences between populations become attributable to differences in ability as well as residual differences in item functioning. This choice of identification lets the random item parameters explain between-population residual variance. Similarly, we can restrict average item discrimination to be constant across populations. Random parameters then explain residual between-population variation in item discrimination. Of course, for both types of parameters, the preferred case is negligible fluctuations between populations, that is, "measurement invariance." However, as described by Verhagen and Fox (2013), when the fluctuations appear to be substantial, they may be explained by background differences (e.g., culture or gender differences) . Besides, it also possible to allow for cross-classified differences in item characteristics, for example, when cross-national and cross-cultural response heterogeneity is present.

1.4 Parameter Estimation

A Bayesian approach to response modeling is a natural way to account for sources of uncertainty in the estimation of the parameters. Also, it leads to estimation procedures that are easy to implement (e.g., Congdon, 2001).

A fully Bayesian approach requires the specification of prior distributions for all model parameters. In Fox (2010), non-informative inverse gamma priors are specified for the variance components, an inverse Wishart prior is specified for the covariance matrix, and vague normal priors are specified for the remaining mean parameters. Subsequently, a Gibbs sampling procedure is used to sample from the joint posterior distribution of all model parameters. More specifically, following Albert (1992), an augmentation scheme is defined to sample latent continuous responses, whereupon the item parameters and multilevel model parameters can be sampled directly from the full conditional distributions given the latent responses (Fox 2010; Fox & Glas, 2001).

Alternatively, several packages in R and WinBUGS are available to estimate the model parameters. Cho and Cohen (2010) developed WinBUGS programs to estimate a mixture multilevel IRT model. For a 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 generally similar results for each of them.

1.5 Model Fit

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

Different multilevel IRT models can be compared for their goodness of fit using the Deviance Information Criterion (DIC); for an introduction, see Cohen & Cho, vol. 2, chap. 18). The criterion 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, $D\left(\hat{\boldsymbol{\Omega}}\right)$ the deviance evaluated at the posterior mean $\hat{\boldsymbol{\Omega}}$, and p_D represents the effective number of parameters defined as 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 becomes

$$\begin{aligned} p\left\{\mathbf{u}; \boldsymbol{\xi}, \boldsymbol{\gamma}, \sigma^2_\theta, \mathbf{T}\right\} &= \prod_j \int_{\boldsymbol{\beta}_j} \left[\prod_{p|j} \int_{\theta_{pj}} \prod_i p\left(u_{pji} \mid \theta_{pj}, \boldsymbol{\xi}_i\right) \right. \\ &\quad \left. p\left(\theta_{pj} \mid \boldsymbol{\beta}_j, \sigma^2_\theta\right) d\theta_{pj} \right] p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right) d\boldsymbol{\beta}_j, \end{aligned} \tag{1.7}$$

such that the fit of random effects are not explicitly expressed in the (marginal) likelihood.

1.6 Empirical Example

Data from the 2003 assessment in 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 original data and results from PISA can be found at *http://pisa2003.acer.edu.au*. Similar to Fox (2010),

the performances of the subpopulation of Dutch students in mathematics was investigated using various selections of background variables. Besides, a random item parameter multilevel IRT model was used to investigate measurement invariance assumptions across Dutch schools.

1.6.1 Data

The performances in mathematics were measured using 84 items. Students received credit for each item they answered correctly. However, although some items were scored with partial credit, for this example all item responses were coded as zero (incorrect) or one (correct). Each student in PISA 2003 was given a test booklet with different combination of clusters of items, with each mathematics item appeared in the same number of test booklets. In total, 3,829 students were sampled from 150 Dutch schools. Students with less than nine responses were not included in the present analysis.

1.6.2 Model Specification

To investigate individual and school differences in student performances, the following unconditional multilevel IRT model was used to analyze the data:

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

where $j=1,\ldots,150$ represent the selected schools. Three levels of modeling were chosen to account for the within-student, between-student, and between-school variability. To account for item parameter variability, random item parameters were assumed. The multilevel IRT model was identified by fixing the mean and variance of the scale at zero and one, respectively.

1.6.3 Results

All model parameters were estimated using MCMC sampling from the posterior distribution, as implemented in the package mlirt[1]. A total of 10,000 MCMC iterations were run, where the first 1,000 iterations were used as burn-in.

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

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

Multilevel parameter estimates can be biased when point estimates are used as a dependent variable. In order to prevent this from happening, the PISA 2003 results were computed using plausible values for the student's math abilities. The plausible values were random draws from the posterior distributions of the ability parameters given the response data. Their use facilitates the computation of standard errors, while allowing for the uncertainty associated with the ability estimates. Fox (2010, chap. 6) showed that the current multilevel IRT parameters estimates and standard deviations were not significantly different from the estimates obtained using plausible values, given the 95% highest posterior density (HPD) intervals.

The between-student and between-school differences in math performance were explained using background variables. At the student level, female, place of birth (Netherlands or foreign), native vs. foreign first language, and the PISA index of economic, social, and cultural status were used as explanatory variables. At the school level, the mean index of economic, social and cultural status for each school was used to explain variability between the average school performances. In Table 1.1, the multilevel IRT model parameter estimates are given in the column labeled MLIRT. It may be concluded that male students performed slightly better than the female students, native speakers performed better than non-native speakers with a migrant background, and students from more advantaged socio-economic backgrounds generally performed better. Besides, the schools' average index of economic, social, and cultural status had a significant positive effect on their average score.

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

	Unconditional 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]

The multilevel IRT analysis were performed assuming invariance of the items

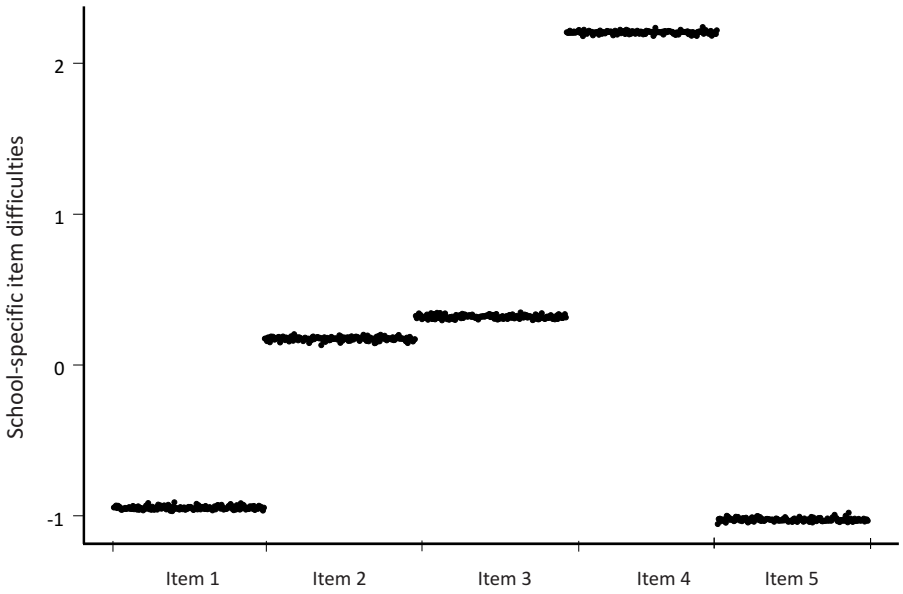

FIGURE 1.1
Random item difficulty estimates of items one to five for the 150 Dutch schools in PISA 2003

across schools. A version of the model with random item parameters with school-specific distributions was used to investigate whether admittance of small deviations in item functioning across schools would lead to a better model fit. Although our preferred model was the measurement invariant multilevel IRT model, this more flexible generalization of it did capture some additional residual variance. Its estimated intra-class correlation coefficient was slightly lower (.43) than for the previous model. However, the item parameter estimates varied hardly across schools; only small variations in item discriminations could be detected. For instance, the first five math items had average difficulty estimates equal to -.94, .17, .32, 2.21, -1.02. In Figure 1.1, the estimated difficulties of these five items are plotted for each school, revealing hardly any differences between schools.

1.7 Discussion

This chapter presented an overview of multilevel IRT modeling. It also reviewed extensions of multilevel models that included student and/or item population distributions. In either case, the clustering of students or items is modeled by an extra hierarchical level in the model. In addition, generalizations to different types of response

data and different types of clustering were presented, and the use of a Bayesian approach with MCMC sampling from the posterior distributions of the parameters was highlighted. The same approach, via posterior predictive assessment, offers the possibility to evaluate model assumptions.

Multilevel response modeling has been shown to be useful in school effectiveness research, where typically differences in responses and abilities within and between schools need to be explored. It also allows us to account for the nested structure of the responses and abilities.

Current multilevel models usually treats the student variable of interest as unidimensional. However, when multiple student abilities must be assumed to produce the observed responses, a multidimensional IRT model can be specified as response model. The specification then requires a multivariate second-level model. The approach has been developed for modeling responses and response times to measure both ability and speed of working (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,* 79–93.

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,* 223–243

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,* 1–41.

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 ed.). 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, 32*, 2988–3005.

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