3

Multilevel IRT Modeling

Akihito Kamata
Department of Educational Methodology, Policy, and Leadership, University of Oregon, Eugene, Oregon

Brandon K. Vaughn
Department of Educational Psychology, University of Texas at Austin, Austin, Texas

3.1 INTRODUCTION

In this chapter, we focus on extending the use of multilevel modeling for psychometric analyses. Such a use of multilevel modeling techniques has been referred to as multilevel measurement modeling (MMM; e.g., Beretvas & Kamata, 2005; Kamata, Bauer, & Miyazaki, 2008). When an MMM considers categorical measurement indicators, such as dichotomously and/or polytomously scored test items, we refer to such a modeling framework as multilevel item response theory (IRT) modeling. Typically, traditional IRT models do not consider a nested structure of the data, such as students nested within schools. However, data in social and behavioral science research frequently have such a nested data structure, especially when data are collected by multistage sampling. The strength of multilevel IRT modeling becomes important when we analyze psychometric data that have such a nested structure. A multilevel IRT model appropriately analyzes data by taking into account both within- and between-cluster variations of the data. Also, since multilevel modeling is essentially an extension of a regression model to multiple levels, the flexibility of multilevel IRT modeling offers the opportunity to incorporate covariates and their interaction effects.

This chapter is organized into three main sections. First, traditional IRT modeling is introduced. Then, a multilevel extension of IRT modeling is presented. In this section, three different modeling frameworks are presented. Lastly, an illustrative data analysis to estimate the variation of differential item functioning (DIF) on a statewide testing program data is presented.
3.2 ITEM RESPONSE THEORY MODELS

Item response theory modeling is a widely utilized class of traditional measurement models. For dichotomously scored test items, there are several well-recognized IRT models, such as the Rasch model, the two-parameter logistic model, and the three-parameter logistic model. For example, the two-parameter logistic model can be written as

\[ p_{ip} = \frac{\exp[\alpha_i \theta_p + \delta_i]}{1 + \exp[\alpha_i \theta_p + \delta_i]}, \]  

(3.1)

where \( \theta_p \) is the ability of examinee \( p \), \( \alpha_i \) is the discrimination power of item \( i \), and \( \delta_i \) is the threshold or location of item \( i \). In IRT applications, the threshold is typically transformed into the difficulty parameter \( \beta_i \) by \( \beta_i = -\delta_i/\alpha_i \), such that the exponential function has a form of \( \alpha_i (\theta_p - \beta_i) \). However, in this chapter we will use the threshold parameter directly for simplicity from a modeling perspective. The metric of \( \theta_p \) and \(-\delta_i/\alpha_i\) are typically in a standardized scale, where 0 is the center of the distribution with a standard deviation of 1. When discrimination power is assumed to be equal for all items in the instrument and constrained to be 1, the model becomes

\[ p_{ip} = \frac{\exp[\theta_p + \delta_i]}{1 + \exp[\theta_p + \delta_i]}, \]  

(3.2)

and is known as the Rasch model. The difference between \( \theta_p \) and \(-\delta_i = \beta_i \) is directly a logit quantity, where \( \theta \) indicates a typical ability or difficulty, respectively. Furthermore, the two-parameter logistic model can be extended to the three-parameter logistic model

\[ p_{ip} = \gamma_i + (1 - \gamma_i) \frac{\exp[\alpha_i \theta_p + \delta_i]}{1 + \exp[\alpha_i \theta_p + \delta_i]}, \]  

(3.3)

where \( \gamma_i \) is the lower asymptote of the logistic curve and known as the pseudo guessing parameter. Under the three-parameter logistic model, we assume a nonzero lower asymptote, indicating a nonzero probability of endorsing an item for examinees with any ability level.

Item response modeling may be extended to polytomously scored items. One widely used model is the Graded Response Model (Samejima, 1969), which utilizes the cumulative logit principle. The model is written as

\[ p_{ijp} = \frac{\exp[\alpha_i \theta_p + \delta_{ij}]}{1 + \exp[\alpha_i \theta_p + \delta_{ij}]} \]  

(3.4)

where \( p_{ijp} \) is the probability for person \( p \) getting the scoring category \( j \) or higher on item \( i \). In this model, \( \delta_{ij} \) is the threshold parameter for the \( j \)th score boundary. As a result, the probability of getting a specific scoring category \( j \) is obtained by \( p_{ijp} = p_{ijp} - p_{(j+1)p} \). For the lowest scoring category \((j = 0)\), \( p_{0ip} = 1 - p_{1ip} \), while \( p_{Mip} = p_{Mip} \) for the highest scoring category \( M \). If \( \delta_{ij} \) is transformed into \( \beta_{ij} = -\delta_{ij}/\alpha_i \), \( \beta_{ij} \) is the category-boundary difficulty for the \( j \)th score boundary. By assuming the discrimination coefficients are equal across all items, it is also sensible to make a one-parameter extension from this model. Another class of IRT models for polytomously scored items is based on the adjacent logit principle. One general form is the generalized partial credit model (Muraki, 1992).
Multilevel IRT Modeling

3.3 Multilevel Item Response Modeling

A multilevel IRT model extends the above mentioned IRT models, such that they consider variations of abilities between group units such as schools, as well as within group units. Accordingly, a multilevel IRT model will distinguish the individual-level abilities and group-level abilities. For example, a multilevel extended two-parameter logistic IRT model for dichotomously scored items could be expressed as

\[
P_{ip} = \frac{\exp \left[ \sum_{j=0}^{x} (\alpha_i \theta_p + \delta_{ij}) \right]}{\sum_{r=0}^{m_i} \exp \left[ \sum_{j=0}^{r} (\alpha_i \theta_p + \delta_{ij}) \right]},
\]

where \( x \) is the target response category for the item, and \( m_i \) is the highest response category for item \( i \) \((j = 0, \ldots, r, \ldots, m)\). Simpler variations of this model include the partial credit model with \( \alpha_i = 1 \) for all \( i \) (Masters, 1982), and the rating scale model with \( \alpha_i = 1 \) for all \( i \) and \( \delta_{ij} = \eta_i + \kappa_j \), where \( \eta_i \) is the item location parameter and \( \kappa_j \) is the step parameter. In the rating scale model, step parameters \( \kappa_j \) are common to all items, indicating distances between step parameters are the same for all items.

3.3.1 Fox and Glas’s Multilevel IRT Modeling

One aspect of multilevel IRT modeling traces back to the development of latent regression model (Vehelst & Eggen, 1989; Zwinderman, 1991, 1997), where observed variables are regressed on the latent variable \( \theta \). Fox and Glas (2001) extended this idea to multilevel linear modeling with two-parameter normal ogive and graded response model as the measurement model. This is a multilevel IRT model due to the nature of the multilevel model being embedded in the IRT framework. In effect, it allows modeling the relationship between observed individual and group characteristics and a latent variable represented by both dichotomous and polytomous items. In Fox and Glas’s formulation, the measurement model is either a two-parameter normal ogive model or graded response model. Additionally, this model describes the structural relationship between the latent variable in the IRT model (ability) \( \xi_g \) that is the amount of deviation from the group mean ability for person \( p \) in group \( g \). This is one of the simplest forms of a multilevel IRT model. However, typical applications of multilevel IRT models involve covariates in the model.

Several different ways to formulate a multilevel IRT model have been presented in the literature. In this section, two approaches, Fox & Glas’s (2001) multilevel IRT framework, and Kamata’s (2001) HGLM approach to multilevel IRT will be presented. We will also describe multilevel structural equation modeling with categorical measurement indicators since both of these approaches can be viewed as special cases of the SEM.
and observed covariates. Thus, the level-2 model is a structural model
\[
\theta_{pg} = \beta_{0g} + \beta_{1g} x_{1pg} + \ldots + \beta_{Qg} x_{Qpg} + \xi_{pg}^{(2)},
\]
(3.7)
where \(\theta\) is the latent variable that represents the trait measured in the measurement (IRT) model, \(x\) are level-2 covariates, \(\beta_{ig}\) are corresponding coefficients, and \(\xi_{pg}^{(2)}\) is the error, where \(\xi_{pg}^{(2)} \sim N(0, \sigma^2)\). Additionally, three-level models can be written as
\[
\begin{align*}
\beta_{0g} &= \gamma_{00} + \gamma_{01} w_{1g} + \ldots + \gamma_{05} w_{5g} + \xi_{0g}^{(3)} \\
\vdots \\
\beta_{Qg} &= \gamma_{Q0} + \gamma_{Q1} w_{1g} + \ldots + \gamma_{Q5} w_{5g} + \xi_{Qg}^{(3)},
\end{align*}
\]
(3.8)
where \(w\) are level-3 covariates, \(\gamma\) are corresponding coefficients, and \(\xi_{0g}^{(3)}, \ldots, \xi_{Qg}^{(3)}\) are level-3 random effects, where \(\xi_{pg}^{(2)} \sim N(0, \Omega)\). If there is no covariate in either levels of the structural models, the structural model is reduced to
\[
\theta_{pg} = \xi_{pg}^{(3)} + \xi_{pg}^{(2)},
\]
(3.9)
since \(\beta_{0g}\) and \(\gamma_{00}\) become the means of \(\xi_{0g}^{(3)}\) and \(\xi_{0g}^{(2)}\), which are 0. This equation demonstrates its equivalency to the general multi-level IRT model equation presented in the previous section (Equation 3.6).

Fox and Glas (2001) and Fox (2005) have implemented a Markov chain Monte Carlo (MCMC) method to estimate the parameters in this model. An \(R\) package for the MCMC called \texttt{mlirt} has been made available for public (Fox, 2007).

### 3.3.2 HGLM Approach

We now focus on the use of hierarchical generalized linear models (HGLMs) for latent variable modeling. The uniqueness of the GLM over general linear models is in the dependent measure. The GLM allows response measures that follow any probability distribution in the exponential family of distributions. Generalized linear models are of great benefit in situations where the response variables follow distributions other than the normal distribution and when variances are not constant. This is of particular interest in IRT as response measures are typically dichotomous or polytomous, discrete, and nonnormal.

The analysis of the GLM incorporates the use of a link since the dependent measure in GLMs may characterize many different types of distributions and thus the relationship between the predictor and the dependent measure may not be linear in nature. Many different link functions exist, yet Table 3.1 shows the most common in research and practice.

The HGLM approach provides a flexible and efficient framework for modeling non-normal data in situations when there may be several sources of error variation. This is accomplished by extending the familiar GLM to include additional random terms in the linear predictor. One special case of HGLMs is generalized linear mixed models (GLMMs), which constrains the additional terms to follow a normal distribution and to have an identity link. However, many HGLMs do not have such restrictions. For example, if the basic GLM is a log-linear model (Poisson

<table>
<thead>
<tr>
<th>Probability Distribution</th>
<th>Link Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Identity</td>
</tr>
<tr>
<td>Binomial/normal cumulative</td>
<td>Logit/probit</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log</td>
</tr>
<tr>
<td>Multinomial</td>
<td>Logit/probit</td>
</tr>
</tbody>
</table>
distribution and log link), a more appropriate assumption for the additional random terms might be a gamma distribution and a log link. Thus, HGLMs bring together a wide range of models under a common approach. Each HGLM is made up of at least two levels in a multilevel model so as to incorporate several sources of error variation. This approach is especially useful in situations involving nested or clustered data. In IRT analysis, this might manifest itself in situations of students nested within schools or individuals nested within families. By considering cluster effects, innovative questions can be considered (e.g., if any differential item functioning, DIF, effects vary from cluster to cluster).

### 3.3.2.1 Modeling IRT as Latent HGLM

Earlier in this chapter, various IRT models were shown. An IRT model can be modeled with a two-level logistic regression where the log-odds (i.e., logit link function) of subject $p$ providing a positive answer to an item $i$ is represented by:

$$
\eta_{ip} = \log \left( \frac{\Phi_{ip}}{1 - \Phi_{ip}} \right) = \theta_p - \beta_i,
$$

(3.10)

where $\Phi_{ip}$ represents the probability that subject $p$ gets item $i$ correct, $\theta_p$ represents the trait level associated with subject $p$ ($\theta_p \sim N(0, \sigma^2)$), stating that $\theta_p$ is normally distributed with 0 mean and the variance of $\sigma^2$), and $\beta_i$ represents the difficulty of item $i$. In this model, $\eta_{ip}$ represents the log-odds of subject $p$ getting item $i$ correct (assuming dichotomous outcomes). This simple IRT model is the Rasch model as detailed earlier. Adding one additional parameter $\alpha_i$ to represent the extent to which item $i$ can discriminate between subjects of different trait levels, the model becomes:

$$
\eta_{ip} = \alpha_i (\theta_p - \beta_i) = \alpha_i \theta_p - \alpha_i \beta_i.
$$

(3.11)

Finally, if a predictor is added to this model in order to provide an explanatory approach, the formulation becomes

$$
\eta_{ip} = \alpha_i \theta_p - \alpha_i \beta_i + \gamma X_p,
$$

(3.12)

where $\gamma$ is the regression coefficient for explanatory variable $X_p$.

For a set of $r$ items, the logit link function can be modeled as a hierarchical two-level logistic model (e.g., Van den Noortgate & De Boeck, 2005):

$$
\eta_{ip} = \log \left( \frac{\Phi_{ip}}{1 - \Phi_{ip}} \right) = \beta_{0p} X_{1ip} + \ldots + \beta_{rp} X_{rip}
$$

$$
+ u_p = \sum_{q=1}^{r} \beta_{qp} X_{qi} + u_p,
$$

(3.13)

where $X_{qi} = 1$ if $q = i$, 0 otherwise, and $u_p \sim N(0, \sigma^2)$. Kamata (2001) parameterized the multilevel logistic model as:

$$
\eta_{ip} = \log \left( \frac{\Phi_{ip}}{1 - \Phi_{ip}} \right)
$$

$$
= \beta_{0p} + \beta_{1p} X_{1ip} + \ldots + \beta_{(r-1)p} X_{(r-1)ip}
$$

$$
= \beta_{0p} + \sum_{q=1}^{r-1} \beta_{qp} X_{qp}.
$$

(3.14)

Each $X_{qp}$ represents the $q$th dummy indicator variable for subject $p$. In order for the design matrix of the model to achieve full rank, one of the items must be dropped from the model or a no-intercept model could be fit. For the case where an item is dropped, for $r$ set of items, only $r - 1$ items are included in the model. The coefficient, $\beta_{0p}$, is interpreted
as the mean effect of the dropped item, and each \( \beta_{qp} \) is interpreted as the effect of the \( q \)th dummy indicator (i.e., item \( i \), for \( i = 1, \ldots, \ r - 1 \)) compared to the reference item. For a particular item \( i \), a value of zero is assigned to \( X_{qip} \) for \( q \neq i \), and a positive one when \( q = i \). This gives a logit for a particular item \( i \), as:

\[
\eta_{ip} = \log \left( \frac{\phi_{ip}}{1-\phi_{ip}} \right) = \beta_{0p} + \beta_{ip},
\]

(3.15)

where \( \beta_{0p} \) is a random effect in which \( \beta_{0p} \sim N(0, \sigma^2_{\beta}) \).

There are a variety of methods to extend this idea to ordinal polytomous outcomes. One popular approach is the formation of a cumulative probability model. For each ordered response \( m (m = 1, \ldots, M) \), a probability of response \( y_i \) on item \( i \) is established for each unique response possibility:

\[
\eta_{mi} = \begin{cases} 
\beta_0 + \beta_1 X_i, & \text{for } m = 1 \\
\beta_0 + \beta_1 X_i + \delta_2 + \ldots + \delta_m, & \text{for } 1 < m \leq M - 1 
\end{cases}
\]

(3.19)

\[
\varphi_m = P(y_i \leq m). 
\]

(3.16)

Defining the probability response model in this manner creates difficulty in formulating a single regression model. Thus, a cumulative probability model is incorporated:

\[
\varphi_m^* = P(y_i \leq m) = \varphi_1 + \varphi_2 + \ldots + \varphi_m. 
\]

(3.17)

A cumulative logit function can be derived using the cumulative probabilities

\[
\eta_m = \log \left( \frac{\varphi_m^*}{1-\varphi_m^*} \right) = \log \left( \frac{P(y_i \leq m)}{P(y_i > m)} \right),
\]

(3.18)

for each ordinal response of \( m = 1, \ldots, M - 1 \). In this model, \( \eta_m \) represents the log-odds of responding at or below category \( m \), versus responding above category \( m \).

A common intercept can be introduced into this model by considering the difference (\( \delta \) ) between the thresholds. The general logit model now becomes

\[
\eta_{mip} = \log \left( \frac{\phi_{mip}}{1-\phi_{mip}} \right) = \begin{cases} 
\beta_{0p} + \beta_{1p} X_{1ip} + \ldots + \beta_{(r-1)p} X_{(r-1)ip}, & \text{for } m = 1 \\
\beta_{0p} + \beta_{1p} X_{1ip} + \ldots + \beta_{(r-1)p} X_{(r-1)ip} + \delta_2 + \ldots + \delta_{mp}, & \text{for } 1 < m \leq M - 1
\end{cases}
\]

(3.20)

This approach can be used to model a two-level HGLM for polytomous items with the IRT perspective mentioned previously. The level-1 (item-level) model for a set of \( r \) items is represented as:
where $\Phi_{mip}^*$ is the cumulative probability as defined above. One random effect, $\beta_{0p}$, is present that represents the expected effect of the reference item for subject $p$. For a particular item $i$, a value of positive one is assigned to $X_{qip}$ when $q = i$, and a value of zero otherwise. For a particular item $q$, this model simplifies to:

$$\eta_{mip} = \beta_{0p} + \sum_{s=2}^{m} \delta_{sj}.$$  

(3.21)

One possible level-2 model (subject-level) with level-2 predictor $X_p$ added to all effects and thresholds is expressed as:

$$\begin{align*}
\beta_{0p} &= \gamma_{00} + \gamma_{01} X_p + u_{0p} \\
\beta_{1p} &= \gamma_{10} + \gamma_{11} X_p \\
\vdots \\
\beta_{(r-1)p} &= \gamma_{(r-1)0} + \gamma_{(r-1)1} X_p \\
\delta_{2p} &= \xi_{20} + \xi_{21q} X_p \\
\vdots \\
\delta_{mp} &= \xi_{m0} + \xi_{m1q} X_p.
\end{align*}$$  

(3.22)

More than one predictor can be incorporated and can be a variety of variables of interest to the researcher that are subject related. For example, in DIF studies, a categorical level-2 predictor of group affiliation (reference versus focal group) can be considered. This modeling can easily be extended to a three-level model. If a third level is added, the level-2 terms can be allowed to vary among clusters of subjects and level-3 predictors (cluster related) can be added to explain the random nature of level-2 terms.

A variety of estimation procedures can be utilized for these HGLM multilevel models. With logistic regression models, estimation procedures have typically incorporated a maximum likelihood method (De Boeck & Wilson, 2004). However, use of this estimation method can prove problematic for multilevel models.

Penalized quasi-likelihood (PQL) estimation was at one time a popular approach. However, this method has been shown to produce negatively biased parameter estimates (Raudenbush, Yang, & Yosef, 2000). Raudenbush et al. (2000) and Yang (1988) suggested a sixth order Laplace (Laplace6) approximation for estimation instead. Current software, such as HLM 6 (Raudenbush, Bryk, & Congdon, 2005), allows for a Laplace6 approximation, but is limited to Bernoulli models of two and three levels. For ordinal models, however, the PQL estimation procedure is still widely used (typically because alternative methods are not widely available in some software packages).

Due to this, some suggest a Bayesian approach as a more flexible option (Johnson & Albert, 1999) and some multilevel software (e.g., MLWin) now have this estimation procedure as an option. Breslow (2003) showed that a MCMC approach is a better choice over PQL for complex problems that involve high dimensional integrals.

Many studies that do approach regular multilevel models from a Bayesian perspective use a probit link function in their formulation (Elrod, 2004; Fox, 2005; Galindo, Vermunt, & Bergsma, 2004; Hoijtink, 2000; Mwalili, Lesaffre, & Declerck, 2005; Qiu, Song, & Tan, 2002). Also popular in certain studies that have considered a Bayesian multilevel approach is the cumulative logit function (Ishwaran, 2000; Ishwaran & Gatsonis, 2000; Lahiri & Gao, 2002; Lunn, Wakefield, & Racine-Poon, 2001). Within this Bayesian framework, MCMC Gibbs sampling estimation procedures are typically used. A variety
of software (e.g., WinBUGS, BRugs for R, MLwiN, etc.) allow for this Gibbs sampling estimation procedure for multilevel models. Although the use of Gibbs sampling has grown in popularity since the advent of powerful personal computers, some psychometric areas still consider Gaussian quadrature points instead for estimation.

Chaimongkol (2005) and Vaughn (2006) both incorporated this approach in estimating random DIF in multilevel models for dichotomous and polytomous items. Vague priors were used in the estimation so that the estimated values would closely mirror those using frequentist methods. In order for the model to be identified, both authors replaced the model parameters with new “adjusted” quantities that were well identified yet did not change the logit of the model.

Although the above mentioned estimation procedures are the most common in practice, there are many others available that might be considered. Goldstein and Rasbash (1992) detail an iterative generalized least squares (IGLS) method for estimation. This approach is sometimes referenced as PQL2 and is incorporated in the computer program MLwiN. Also, as mentioned above, Gaussian quadrature estimation is a popular choice in other software (e.g., Sabre, Stata, and GLLAMM).

### 3.3.3 Multilevel SEM Approach

A more general framework for a multilevel IRT modeling is a two-level structural equation model with categorical indicators. The two-level SEM assumes that multiple individuals are sampled from each of many groups in the population (see Muthén and Asparouhov, Chapter 2 of this book).

The two-level factor model with categorical indicators can be written as

\[
\mathbf{y}_{pg}^* = \Lambda_w \mathbf{\theta}_{pg} + \mathbf{\epsilon}_{pg},
\]

which represents a linear regression of the vector of I unobserved latent response variables \( \mathbf{y}_{pg}^* \) on the latent variables \( \mathbf{\theta}_{pg} \) for person \( p \) in group \( g \). The latent response variables \( \mathbf{y}_{pg}^* \) is an \( I \times 1 \) vector of latent response scores to \( I \) items in the test, and \( \mathbf{\theta}_{pg} \) is a \( K \times 1 \) vector of factor scores (abilities) for \( K \) latent factors. As a result, \( \Lambda_w \) are factor loadings (\( I \times K \) matrix), where the \( W \) subscript indicates “within-groups,” and \( \mathbf{\epsilon}_{pg} \) are residuals (\( I \times 1 \) vector). In a unidimensional IRT application, for example, \( K = 1 \), and both \( \Lambda_w \) and \( \mathbf{\epsilon}_{pg} \) are \( I \times 1 \) vectors. Observed dichotomous response \( y_{ipg} \) is defined such that

\[
y_{ipg} = 1, \text{ if } y_{ipg}^* \geq \tau_i, \text{ and } \]
\[
y_{ipg} = 0, \text{ if } y_{ipg}^* < \tau_i. \tag{3.24}
\]

Here, \( \tau_i \) is the threshold for item \( i \). Within groups, the latent factors are assumed to be distributed with mean vector \( \alpha \) and covariance matrix \( \Psi_w \). Similarly, for polytomously scored items with scoring categories ranging from 0 to \( M \),

\[
y_{ipg} = M, \text{ if } y_{ipg}^* \geq \tau_{iM},
\]
\[
y_{ipg} = M - 1, \text{ if } \tau_{(i-1)M} \leq y_{ipg}^* \leq \tau_{iM},
\]
\[
\vdots
\]
\[
y_{ipg} = 1, \text{ if } \tau_{i1} \leq y_{ipg}^* \leq \tau_{i2}, \text{ and } \]
\[
y_{ipg} = 0, \text{ if } y_{ipg}^* < \tau_{i1}.
\]

The residuals \( \mathbf{\epsilon}_{pg} \) are assumed to be distributed with means of zero and covariance
Residuals are independent from each other according to the local independence assumption of IRT models, resulting in a diagonal $\Sigma_w$ matrix. If errors are distributed as the logistic distribution, the model is known as the logistic model, and this will provide the basis of equivalency to logistic item response models. If we have $\theta_{pg}$ as a $1 \times 1$ scalar (i.e., only one latent trait) and $M=2$ (i.e., dichotomously scored items), the model is equivalent to the 2PL IRT model. On the other hand, if residuals are normally distributed, the model is known as the normal ogive model. One important assumption with this approach is that these covariance matrices are homogeneous across all groups, which will result in identical covariance structures between groups. Accordingly, for group $j$, the within-group covariance matrix is

$$V(y^*)_w = \Lambda_w \psi_w \Lambda'_w + \Sigma_w, \quad (3.26)$$

which is essentially the same for the single-level SEM, except the $W$ subscript for each quantity in the equation. On the other hand, the structural model of the two-level SEM can be written as

$$\theta_{pg} = \alpha_{pg} + B_{pg} \theta_{pg} + \Gamma_{pg} x_{pg} + \xi_{pg}, \quad (3.27)$$

where latent factors are regressed on other latent factors and some observed covariates $x$. The intercepts are given by $\alpha_{pg}$, slopes for latent predictors are $B_{pg}$, and slopes for observed covariates are $\Gamma_{pg}$. The residuals are assumed to be normally distributed with means of zero and $K \times K$ covariance matrix $\psi$. If no latent variable is specified as a predictor in the model, the intercepts, $\alpha_{pg}$, are simply factor means, which are typically constrained to be 0.

Due to nested data structure, the multilevel SEM imposes an additional between-group level factor structure on the covariance matrix (e.g., Ansari, Jedidi, & Dube, 2002; Goldstein & McDonald, 1988; McDonald & Goldstein, 1989; Muthén, 1994; Muthén & Satorra, 1995). The resulting covariance structure at between-group level is

$$V(y^*|x)_B = \Lambda_B \psi_B \Lambda'_B + \Sigma_B, \quad \text{and}$$

$$V(y^*|x)_B = \Lambda_B (I - B_B)^{-1} \psi_B (I - B_B)^{-1} \Lambda'_B + \Sigma_B. \quad \text{(3.28)}$$

Here, the structure for the within- and between-groups covariance matrices are very similar (Equations 3.25 and 3.27). However, the parameter estimates and the factor structure of the model can be different between the two parts of the model as indicated by different subscripts ($W$ vs. $B$).

Traditionally, parameter estimation for this type of model has relied on the weighted least squares methods with a tetrachoric or polychoric correlation matrix, which differs from the IRT estimation tradition, where a full information maximum likelihood has been a common approach. Also, the scaling of parameters will be different from the parameter scale of IRT if the weighted least square is employed, which requires appropriate transformation of parameters (e.g., Kamata et al., 2008). More recently, a true full information maximum likelihood estimator has become available in several general SEM software programs, which is consistent with the IRT tradition. Also, the MCMC has been shown to be effective for this type of model, especially when the
number of random effects becomes large (e.g., Chaimongkol, 2005; Fox, 2005; Fox & Glas, 2001; Vaughn, 2006).

3.4 ILLUSTRATIVE DATA ANALYSIS

Data used in the following illustrative analyses were sampled from the 2005 administration of mathematics assessment for eighth graders in a statewide testing program in one particular state in the United States. The mathematics test consisted of 40 items based on five subscales of skills, including: Number Sense, Measurement, Geometry, Algebraic Thinking, and Data Analysis. Here, only one subscale, Data Analysis, was used for illustrative data analysis. With the Data Analysis subscale, items measured various skills to use and interpret data through mean, median, and probability, as well as the use of a Venn diagram. Among nine items in this subscale, eight items were scored dichotomously and one item was scored polytomously with five ordered scoring categories. All nonresponded items were scored as 0. The sample of examinees included a total of 11,220 examinees from 30 schools.

By using this data set, modeling a random differential item functioning (RDIF) is demonstrated. An RDIF is a differential item functioning (DIF) that is treated as a random effect. To be more specific, we consider that the magnitude of DIF for a particular test item to vary across schools. Here, we evaluate the RDIF between English language learners (ELL) and students in standard curricula. All model parameters are estimated by full-information maximum likelihood with adaptive numerical integration. *Mplus* syntax for the three analyses presented here are provided in the Appendix.

First, a DIF detection model was fitted under the multilevel CFA with a covariate. The model can be written

\[
y'_{ipg} = \lambda_i \theta_{pg}^{(2)} + \beta_i G_{pg} + \varepsilon_{ipg} \tag{3.29}
\]

with structural models

\[
\theta_{pg}^{(2)} = \beta_0 + \beta_0 G_{pg} + \xi_{pg}^{(2)}, \text{ and}
\]

\[
\beta_0 = \xi_{p(3)}.
\tag{3.30}
\]

When these measurement and structural models are combined, the model is

\[
y'_{ipg} = \lambda_i (\xi_{pg}^{(2)} + \xi_{p(3)} + \beta_0 G_{pg}) + \beta_i G_{pg} + \varepsilon_{ipg},
\tag{3.31}
\]

where \(G_{pg}\) is the group indicator (1 = ELL, 0 = Standard Curricula). Consequently, \(\beta_0\) indicates the difference between ELL and Standard Curricula students in their ability, and \(\beta_i\) is the DIF magnitude for item \(i\). This is essentially the same as the MIMIC approach to DIF detection (e.g., Finch, 2005) with an additional cluster level. Since this model is not identified, it was constrained with the magnitude of DIF for the first item being zero. This constraint was based on our initial data screening, including single-item DIF analysis for all items on the entire mathematics test. Therefore, we are quite confident that this constraint is not too far from reality, and that estimated DIF magnitudes for the remaining items can be interpreted as the magnitude of DIF. However, one might be more on the conservative side to interpret \(\beta_i\) as the difference between the first item and the \(i\)th item in their DIF magnitudes. Note that this limitation in scaling the DIF parameter is not unique.
to this current approach; it is an inherent problem in DIF analysis in general. Bolt, Hare, and Newmann (2007) and Penfield and Camilli (2007) provide more detailed discussions on this matter. Furthermore, the mean of the within-level and between-level latent factor were constrained to zero and the variance of the within-level latent factor was constrained to be one. Although it is not required for identification, item discriminations (factor loadings) were constrained to be the same for within- and between-levels. Results are summarized in Table 3.2. Values in the table without parenthesis indicate estimates of parameters and the values inside parentheses are standard errors of estimates.

The second model assumed that one of the DIF magnitudes was a random effect, where the assumption was that DIF varied across schools. This effect was modeled as another random effect $\xi^{(3)}_{ig}$ in an additional structural model

$$\beta_i = \gamma_{0i} + \xi^{(3)}_{ig}. \quad (3.32)$$

Here, the DIF for the $i$th item is expressed as the sum of the mean of DIF for the item across schools $\gamma_{0i}$ and the random effect $\xi^{(3)}_{ig}$. The variance of this random effect is our
main interest. The combined model can be written as

\[ y_{ipg}^* = \lambda_i (\zeta_i^{(2)} + \xi_i^{(3)} + \beta_0 G_{pg}) + (\gamma_{0i} + \xi_i^{(3)})G_{pg} + \epsilon_{ipg}, \quad (3.33) \]

In the data analysis, the DIF of item 7 was treated as a random effect. The results are shown in Table 3.3. Again, values in the table without parenthesis indicate estimates of parameters and the values inside parentheses are standard errors of estimates.

### Table 3.3

Results of Data Analysis with RDIF Model

<table>
<thead>
<tr>
<th>Item</th>
<th>Loading</th>
<th>Thresholds</th>
<th>DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.931</td>
<td>−.710</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(.035)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.446</td>
<td>−2.893</td>
<td>−0.225</td>
</tr>
<tr>
<td></td>
<td>(0.061)</td>
<td>(0.086)</td>
<td>(.118)</td>
</tr>
<tr>
<td>3</td>
<td>0.848</td>
<td>−1.186</td>
<td>−0.157</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.040)</td>
<td>(.113)</td>
</tr>
<tr>
<td>4</td>
<td>1.142</td>
<td>−2.628: −1.717: 1.113: 2.131</td>
<td>−0.808</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(.056) (.050) (.047) (.040)</td>
<td>(.096)</td>
</tr>
<tr>
<td>5</td>
<td>0.891</td>
<td>0.220</td>
<td>−0.146</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.045)</td>
<td>(.100)</td>
</tr>
<tr>
<td>6</td>
<td>0.716</td>
<td>−0.102</td>
<td>−0.121</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.035)</td>
<td>(.093)</td>
</tr>
<tr>
<td>7</td>
<td>1.300</td>
<td>−0.707</td>
<td>See the mean of random effect for item 7</td>
</tr>
<tr>
<td></td>
<td>(0.063)</td>
<td>(.061)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.237</td>
<td>−0.354</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
<td>(0.045)</td>
<td>(.119)</td>
</tr>
<tr>
<td>9</td>
<td>0.893</td>
<td>−0.991</td>
<td>−0.117</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.041)</td>
<td>(.089)</td>
</tr>
</tbody>
</table>

**b. Random Effects**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-level latent factor</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Between-level latent factor</td>
<td>0.000</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.018)</td>
</tr>
<tr>
<td>DIF for item 7</td>
<td>−0.295</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>(.136)</td>
<td>(.150)</td>
</tr>
</tbody>
</table>

Note that the variance of DIF for item 7 was found to be 0.419, which is equivalent to \( SD = 0.647 \). Assuming a normal distribution of DIF magnitudes across schools, it can be interpreted that the range of DIF for the middle 68% of schools is nearly 1.30, which is quite large around a reasonable logit value.

Next, one covariate was entered into the model in an attempt to explain the variation of the DIF for item 7. The covariate used here is the proportion of limited English proficient (LEP) students in each school. The variable was scaled such as
10% = 1 unit. Some descriptive statistics were as follows; N = 30, minimum = .34, maximum = 4.73, mean = 1.19, SD = .90. Since one part of the structural model is expanded to \( \beta_i = \gamma_{0i} + \gamma_{1i}(LEP_P)_g + \xi^{(3)}_g \), the model is now written

\[ y_{pg}^* = \lambda_i(\zeta_{pg}^{(2)} + \xi^{(3)}_g + \beta_g G_{pg}) + \left[ \gamma_{0i} + \gamma_{1i}(LEP_P)_g + \xi^{(3)}_i \right]G_{pg} + \epsilon_{ipg}. \tag{3.34} \]

The results are summarized in Table 3.4. Again, values in the table without parenthesis indicate estimates of parameters and the values inside parentheses are standard errors of estimates. The mean of the DIF of item 7 is now expressed as a linear function of LEP_P.

Note that the variance of DIF for item 7 was .419, which is equivalent to \( SD = .647 \) in the previous model. When a school-level covariate LEP_P is added to the model, this variance was reduced to .224 (65% reduction of the variance). The estimated coefficient for the covariate is .425 (SE = .086), indicating positive relationship between the proportion of ELL students and the DIF magnitude.

### Table 3.4

Results of Data Analysis for RDIF Model with a Covariate

<table>
<thead>
<tr>
<th>Item</th>
<th>Loading</th>
<th>Thresholds</th>
<th>DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.930</td>
<td>-.710</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(.029)</td>
<td>(.035)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.446</td>
<td>-.892</td>
<td>-225</td>
</tr>
<tr>
<td></td>
<td>(.061)</td>
<td>(.086)</td>
<td>(.117)</td>
</tr>
<tr>
<td>3</td>
<td>0.848</td>
<td>-.186</td>
<td>-157</td>
</tr>
<tr>
<td></td>
<td>(.036)</td>
<td>(.040)</td>
<td>(.112)</td>
</tr>
<tr>
<td>4</td>
<td>1.142</td>
<td>-.628</td>
<td>-147</td>
</tr>
<tr>
<td></td>
<td>(.045)</td>
<td>(.056)</td>
<td>(.100)</td>
</tr>
<tr>
<td>5</td>
<td>0.891</td>
<td>-.220</td>
<td>-122</td>
</tr>
<tr>
<td></td>
<td>(.040)</td>
<td>(.045)</td>
<td>(.092)</td>
</tr>
<tr>
<td>6</td>
<td>0.716</td>
<td>-.102</td>
<td>-118</td>
</tr>
<tr>
<td></td>
<td>(.035)</td>
<td>(.035)</td>
<td>(.119)</td>
</tr>
<tr>
<td>7</td>
<td>1.303</td>
<td>-.707</td>
<td>See the mean of random effect for item 7</td>
</tr>
<tr>
<td></td>
<td>(.063)</td>
<td>(.061)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.236</td>
<td>-.354</td>
<td>.014</td>
</tr>
<tr>
<td></td>
<td>(.054)</td>
<td>(.045)</td>
<td>(.109)</td>
</tr>
<tr>
<td>9</td>
<td>0.892</td>
<td>-.991</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.033)</td>
<td>(.041)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-level latent factor</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Between-level latent factor</td>
<td>0.000</td>
<td>.090</td>
</tr>
<tr>
<td>DIF for item 7</td>
<td>-.836 + .425 (LEP_P)</td>
<td>.224</td>
</tr>
<tr>
<td></td>
<td>(.190)</td>
<td>(.086)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.102)</td>
</tr>
</tbody>
</table>
Since the intercept is negative – .836, the interpretation is that schools with a very small proportion of ELL students had a larger disadvantage for ELL students for this item. As the proportion of ELL increases, schools had smaller disadvantages for this item. For schools with about 20% ELL students, DIF is predicted to be near zero, by substituting \( LEP_P = 2 \) into \( -.836 + .425(LEP_P) \).

### 3.5 CONCLUSIONS

Multilevel IRT modeling as an extension of multilevel modeling was discussed in this chapter. Several different modeling frameworks were introduced. As a practical application of the multilevel IRT modeling, estimation of random DIF (RDIF) was demonstrated with a data set sampled from a statewide testing program in the United States. As mentioned earlier, traditional IRT models typically do not consider a nested structure of the data. However, data in social and behavioral science research frequently have such a nested data structure, especially when data are collected by multistage sampling. As demonstrated in this chapter, the strength of multilevel IRT modeling becomes important when we analyze psychometric data that have such a nested structure. A multilevel IRT model appropriately analyzes data by taking into account both within- and between-cluster variations of the data. Also, since multilevel modeling is an extension of a regression model to multiple levels, the flexibility of multilevel IRT modeling offers the opportunity to incorporate previous approaches and techniques, as demonstrated in this chapter.

### REFERENCES


APPENDIX

Mplus Syntax for the Data Analysis Illustrations

1. DIF detections with multilevel IRT model

TITLE: 2PL DIF detection model
DATA: FILE IS math_lep4.dat;
VARIABLE: NAMES ARE sch u1-u9 lep lep_p;
CATEGORICAL ARE u1-u9;
Cluster = sch;
within = lep;
usevariables = sch u1-u9 lep;
ANALYSIS: Type = Twolevel Random;
MODEL: %Within%
  f BY u1* (1)
  u2 (2)
  u3 (3)
  u4 (4)
  u5 (5)
  u6 (6)
  u7 (7)
  u8 (8)
  u9 (9);
  [f@0];
  f@1;
  f on lep;
  u1 on lep@0;
  u2 on lep;
  u3 on lep;
  u4 on lep;
  u5 on lep;
  u6 on lep;
  u7 on lep;
  u8 on lep;
  u9 on lep;

%between%
  fb BY u1* (1)
  u2 (2)
  u3 (3)
  u4 (4)
  u5 (5)
  u6 (6)
  u7 (7)
  u8 (8)
  u9 (9);

2. RDIF detection for item 7 with multilevel 2-PL IRT model

TITLE: 2PL RDIF model
DATA: FILE IS math_lep4.dat;
VARIABLE: NAMES ARE sch u1-u9 lep lep_p;
CATEGORICAL ARE u1-u9;
Cluster = sch;
within = lep;
usevariables = sch u1-u9 lep;
ANALYSIS: Type = Twolevel Random;
MODEL: %Within%
  f BY u1* (1)
  u2 (2)
  u3 (3)
  u4 (4)
  u5 (5)
  u6 (6)
  u7 (7)
  u8 (8)
  u9 (9);
  [f@0];
  f@1;
  f on lep;
  u1 on lep@0;
  u2 on lep;
  u3 on lep;
  u4 on lep;
  u5 on lep;
  u6 on lep;
  u7 on lep;
  u8 on lep;
  u9 on lep;

%between%
  fb BY u1* (1)
  u2 (2)
  u3 (3)
  u4 (4)
  u5 (5)
  s7 | u7 on lep;
  u8 on lep;
  u9 on lep;
3. RDIF explanation with a school-level covariate for item 7 with multilevel 2-PL IRT model

**TITLE:** 2PL RDIF model with a school-level covariate

**DATA:** FILE IS math_lep4.dat;

**VARIABLE:** NAMES ARE sch u1-u9 lep lep_p;

CATEGORICAL ARE u1-u9;
Cluster = sch;
within = lep;
between = lep_p;

**ANALYSIS:** Type = Twolevel Random;

**MODEL:** %Within%
  f BY u1* (1)
  u2 (2)
  u3 (3)
  u4 (4)
  [fb@0];
  s7 on lep_p;

  u5 (5)
  u6 (6)
  u7 (7)
  u8 (8)
  u9 (9);
  [f@0];
  f@1;
  f on lep;
  u1 on lep@0;
  u2 on lep;
  u3 on lep;
  u4 on lep;
  u5 on lep;
  u6 on lep;
  s7 | u7 on lep;
  u8 on lep;
  u9 on lep;

%between%
  fb BY u1* (1)
  u2 (2)
  u3 (3)
  u4 (4)
  u5 (5)
  u6 (6)
  u7 (7)
  u8 (8)
  u9 (9);
  [fb@0];
  s7 on lep_p;