Multilevel Modeling of Educational Data

Edited by

Ann A. O'Connell
Ohio State University

and

D. Betsy McCoach
University of Connecticut

IAP

INFORMATION AGE PUBLISHING, INC.
Charlotte, NC • www.infoagepub.com

CHAPTER 10

MULTILEVEL MEASUREMENT MODELING

Akihito Kamata, Daniel J. Bauer, and Yasuo Miyazaki

Multilevel modeling can be utilized for psychometric analyses, and such a use of multilevel modeling techniques is referred to as multilevel measurement modeling (MMM) (e.g., Beretvas & Kamata, 2005). Typically, traditional psychometric models, including classical test theory (CTT) and item response theory (IRT) models, do not consider a nested structure of the data, such as students nested within schools. However, data in educational research frequently have such a nested data structure, especially when data are collected by multi-stage sampling. The strength of MMM becomes important when we analyze psychometric data that have such a nested structure. MMM 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 MMM offers the opportunity to incorporate covariates and interaction effects. As discussed in previous chapters of this book, another advantage of a multilevel approach is that it can accommodate unbalanced data, using all of the available information in the data.

This chapter is organized into six main parts. First, a brief introduction to traditional measurement models is provided. Second, MMM for continu-

Multilevel Modeling of Educational Data, pages 345–388

ous test components using Hierarchical Linear Models (HLM) is demonstrated. Third, MMM for dichotomously-scored test components using Hierarchical Generalized Linear Models (HGLM) is introduced. Fourth, MMM with covariates is demonstrated for the HGLM approach. Fifth, limitations of HLM/HGLM approaches are discussed. Lastly, an alternative multilevel structural equation modeling (SEM) approach is introduced. Throughout the chapter, a data set from a state-wide testing program is analyzed for illustration purposes.

MEASUREMENT MODELS—TRADITIONAL PERSPECTIVES

Since it is essential to understand measurement models from traditional perspectives, this section briefly reviews classical test theory and item response theory.

Classical Test Theory

In classical test theory (CTT; e.g., Crocker & Algina, 1986; Lord & Novick, 1968; Traub, 1994), an observed score on a test component (e.g., item) is the sum of the underlying true score plus error of measurement. For an observed score on test component (e.g., item) i (X_i),

$$X_i = T_i + E_i, \tag{10.1}$$

where T_i is the true score and E_i is the error score of the test component. Statistically, T_i is the expected value of X_i. On the other hand, E_i is a random variable with mean = 0 and unknown variance, var(E_i).

Depending on the assumptions made about the true and error scores, there are several types of CTT models. These are (strictly) parallel, essentially parallel, (strictly) tau-equivalent, essentially tau-equivalent, and congeneric measures, in (approximately) descending order of strictness of the assumptions. The minimum requirements for the least strict condition (congeneric condition) are that (a) all the test components (e.g., items) measure the same construct; (b) the true scores (T_i) and error scores (E_i) are uncorrelated; and (c) error scores (E_i) are uncorrelated across test components. Other measures need to satisfy these three requirements, in addition to other requirements for T_i and E_i, as described below.

If $T_i = T_{i'}$ and var(E_i) = var($E_{i'}$) for any test components (e.g., items) i and i' ($i \neq i'$), the test components are said to be strictly parallel. If the error variances are the same across test components (e.g., items) but the true scores between test components are different only by a constant ($T_i = T_{i'} + a_{ii'}$ for components i and i' [$i \neq i'$]), the components are said to be essentially parallel. When the true scores differ by only a constant, the implication is that the variances of the true scores across examinees remain the same across test components. More practically, the essentially parallel condition implies that the difference in true scores is the difference in their test component difficulties (e.g., item difficulties), while the reliabilities of test components are equal. If $T_i = T_{i'}$ for all i and i' ($i \neq i'$) but var(E_i) are not the same, the test components are strictly tau-equivalent. This is a situation where the difficulties of test components (e.g., items) are the same while the reliabilities of test components are different. If $T_i = T_{i'} + a_{ii'}$ for all i and i' ($i \neq i'$), where $a_{ii'}$ is a constant, and var(E_i) are not the same, the test components are essentially tau-equivalent. This is a situation where the difficulties of test components (e.g., items difficulties) are not the same and the reliabilities of test components are not equal. Under the parallel and tau-equivalent conditions (including their essential conditions), true scores of test components are perfectly correlated. From a factor analytic perspective, parallel and tau-equivalent conditions are represented by homogeneous factor loadings for test components, indicating equal test component (e.g., item) discriminations. Finally, when $T_i \neq T_{i'}$, which implies that var(T_i) $\neq$ var($T_{i'}$) for any test components (e.g., items) i and i' ($i \neq i'$), and var(E_i) $\neq$ var($E_{i'}$), test components are said to be congeneric. Under the congeneric condition, true scores of test components do not have unit correlations. From a factor analytic perspective, the congeneric condition is represented by heterogeneous factor loadings for test components, indicating non-equal test component (e.g., item) discriminations. Assumptions for the five test conditions are summarized in Table 10.1. For more detailed discussions on the differences in the assumptions, see Traub (1994), and Miyazaki (2005).

One important characteristic of the classical test theory (CTT) model is that the observed variables X_i are treated as continuous, even when test components (e.g., items) are scored dichotomously or polytomously. Also,

TABLE 10.1 Assumptions of the Five CTT Conditions

	$T_i = T_{i'}$?	Var(E_i) = Var($E_{i'}$) ?	Correlation between true scores
(Strictly) Parallel	Yes	Yes	1
Essentially parallel	No, but $T_i = T_{i'} + a_{ii'}$	Yes	1
(Strictly) Tau-equivalent	Yes	No	1
Essentially tau-equivalent	No, but $T_i = T_{i'} + a_{ii'}$	No	1
Congeneric	No	No	Not 1

well-known Cronbach's alpha (an index of internal consistency and measurement reliability under a single administration of a test) is derived based on the CTT framework.

Item Response Theory

Item response theory (IRT) models represent another class of measurement models. For dichotomously scored test items, there are several well-recognized IRT models, such as the Rasch model, the 2-parameter logistic model (2PL), and the 3-parameter logistic model (3PL). The 2PL model can be written as

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_i(\theta_j-\beta_i), \tag{10.2}$$

where θ_j is the ability of an examinee j, α_i is the discrimination of item i, and β_i is the difficulty of item i. The metric of θ_j and β_i are typically in a standardized scale, where 0 is the center of the distribution with a standard deviation of 1. When the discrimination is equal for all items in the test and it is constrained to be 1, the model becomes

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\theta_j-\beta_i \tag{10.3}$$

and is known as the Rasch model. Since the difference between θ_j and β_i is directly a logit quantity, the metric of θ_j and β_i are typically in the logit scale, where 0 indicates a typical ability or difficulty.

The most prominent difference between the CTT and the IRT models is that IRT models treat item response data as a categorical variable rather than as a continuous variable. On the other hand, there is a link between IRT models and CTT models. For example, assumptions about error scores and true scores are similar between 2PL and the congeneric condition and between the Rasch model and essentially tau-equivalent condition (see Miyazaki, 2005). Also, 2PL and the Rasch models are special cases of a factor analytic model (see Kamata & Bauer, 2008; Takane & de Leeuw, 1987).

MULTILEVEL MEASUREMENT MODEL BY HLM AND HGLM APPROACHES

This section presents multilevel measurement models with continuous and dichotomous measurement indicators (test components). When a measurement model is formulated by a linear multilevel model such as HLM (Raudenbush & Bryk, 2002) with test components or items (measurement indicators) that are continuous variables, it corresponds to a classical test theory model. Alternatively, when a measurement model is formulated by a generalized linear multilevel model such as HGLM (Raudenbush & Bryk) with test components or items (measurement indicators) that are categorical variables, the model corresponds to an item response theory model, such as the Rasch model. Whereas CTT and IRT models implicitly assume simple random sampling, multilevel measurement modeling can accommodate data collected by a complex sampling method, such as multistage sampling. Multilevel modeling can accommodate CTT or IRT models into its framework, and thus, it can be used to conduct psychometric analyses for clustered data, such as measures collected from children clustered within classrooms or schools.

Illustrative Data[1]

The example data set includes the fourth-grade mathematics assessment from a statewide testing program in the United States. A total of 3,312 students (N) from 30 schools (K) were sampled, where schools first were selected randomly from the population of schools. For the examples presented here, we use three of the five subscales from the mathematics assessment (eight measurement items, seven algebraic thinking items, and seven data analysis/probability items). Because it is assumed that the subscales measure a single construct, mathematics proficiency, the scale score for the whole test was created simply as the sum of the three subscale scores. All items in these three subscales were scored dichotomously (correct or incorrect). For our first example, illustrating a MMM with continuous measurement indicators, the scale score for each subscale was obtained by computing the proportion of items correct for each subscale. Proportion scores, rather than total scores, were used because each subscale had a different numbers of items. Further, the proportion score for each subscale was multiplied by 10 (this was done solely for the purpose of working with larger numbers that would be easier to present and interpret). Descriptive statistics are provided in Table 10.2. These subscale scores were treated as three measurement indicators (test components) in our first illustration using HLM. After introducing the HGLM formulation, the individual items are used instead as measurement indicators. Therefore, 22 measurement indicators were used for the HGLM analyses.

HLM Approach for the Multilevel Measurement Modeling

In the HLM approach, scores of the test components (subscales, in this illustrative analysis) are treated as continuous variables. Without consider-

TABLE 10.2 Descriptive Statistics for the Sample Data

a. Student level (N = 3312)

Subscale	Minimum	Maximum	Mean	SD
1. Measurement	0	10	5.605	2.510
2. Algebraic thinking	0	10	4.874	2.555
3. Data analysis/probability	0	10	5.556	2.464
Total	0	30	16.035	6.355

b. School level (K = 30)

	Minimum	Maximum	Mean	SD
Number of students	20	228	110.40	47.544
School mean subscale 1 (measurement)	3.45	7.40	5.554	0.935
School mean subscale 2 (algebraic thinking)	3.37	6.85	4.793	0.901
School mean subscale 3 (data analysis/probability)	3.82	7.89	5.513	0.920
School mean total	10.72	21.69	15.860	2.649

ing the nesting of students within schools, the conventional coefficient alpha for the total score of the three subscales was .798 (obtained by the SPSS Reliability procedure). The corresponding standard error of measurement for the total score was estimated as $6.355 \times \sqrt{1-.798} = 2.856$.[2]

Many problems may arise as a consequence of ignoring nested data structures when making statistical inferences. These negative effects are well documented in multilevel modeling textbooks, such as Hox (2005) and Raudenbush and Bryk (2002). However, few studies have focused specifically on the effect of ignoring nested data structures in psychometric analyses. Among them, Raudenbush, Rowan, and Kang (1991) demonstrated that coefficient alpha is inherently ambiguous if nesting is ignored in school-based studies because it measures neither the reliability of school-level measures nor the reliability of student-level measures. Instead, there are actually two types of internal consistency that have clear interpretability in a multilevel design: one at the student level and another at the school level. In the following examples, we will demonstrate how to obtain these two measures of internal consistency, provide their interpretations, and describe how they compare to the conventional coefficient alpha.

To take into account group membership, a multilevel measurement model can be formulated as a three-level hierarchical linear model. We first consider a standard univariate three-level HLM, where the level-one measurement errors are assumed to be homogeneous across groups. This

assumption is equivalent to the essentially parallel condition described earlier, where it is assumed that the subtest true scores differ only in their means and not in their dispersion; that is, scores differ only according to the difficulty of the subtest, and error variances are equal for all subtests. As mentioned earlier, discriminations are assumed equal for the three subtests under this condition.

The level-one model can be written as

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} D2_{ijk} + \pi_{2jk} D3_{ijk} + \varepsilon_{ijk}, \tag{10.4a}$$

where Y_{ijk} is the score on subscale i for student j in school k ($i = 1,\ldots, 3$; $j = 1,\ldots, J_k$ where J_k is the number of students in school k; $k = 1,\ldots, 30$), $D2_{ijk}$ is an indicator taking on a value of 1 if the response i for student j in school k belongs to subscale 2 (algebraic thinking) and 0 if not, and $D3_{ijk}$ is an indicator taking on a value of 1 if the response i belongs to subscale 3 (data analysis/probability) and 0 if not.[3] In this setup, subscale 1 serves as the reference subscale. The intercept, π_{0jk} is the true score of subscale 1 (measurement) for student j in school k. The remaining coefficients are interpreted relative to the reference subscale: π_{1jk} is the true score difference between subscale 2 and subscale 1, and π_{2jk} is the true score difference between subscale 3 and subscale 1. Therefore, the true score of subscale 2 for student j in school k is $\pi_{0jk} + \pi_{1jk}$, and the true score of subscale 3 for student j in school k is $\pi_{0jk} + \pi_{2jk}$. Furthermore, the measurement errors, ε_{ijk}, are assumed to be normally distributed with a mean of zero and constant variance, σ^2. These measurement errors also are assumed to be independent of one another. The latter assumption is equivalent to the assumption of uncorrelated errors in classical test theory.[4] The single value for the error variance, σ^2, represents the homogeneous error variance across subscales assumed for parallel measures.

The level-two model is the student-level model, where π_{0jk}, π_{1jk}, and π_{2jk} are treated as outcome variables. In order to reflect the proposition that all three of the subscales measure a common construct, only the intercept π_{0jk} is treated as random. Accordingly, the level-two equations are

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + r_{0jk} \\ \pi_{1jk} &= \beta_{10k} \\ \pi_{2jk} &= \beta_{20k} \end{aligned}, \tag{10.4b}$$

where β_{00k} is the mean subscale-1 mathematics score for school k, β_{10k} is the mean difference between subscale-1 and subscale-2 mathematics scores for school k, and β_{20k} is the mean difference between subscale-1 and subscale-3 mathematics scores for school k. The random effect, r_{0jk}, is interpreted as

mathematics proficiency level for student j in school k and as independent and normally distributed with mean of 0 and variance τ_π.

The level-three model is the school-level model. Only the intercept, β_{00k}, was assumed to vary randomly across schools. Therefore, the level-three equations are

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + u_{00k} \\ \beta_{10j} &= \gamma_{100} \\ \beta_{20j} &= \gamma_{200} \end{aligned} \quad , \qquad (10.4c)$$

where γ_{000} is the grand mean of subscale-1 score; γ_{100} is the grand mean difference between subscales 1 and 2, γ_{200} is the grand mean difference between subscales 1 and 3. The random effect, u_{00k}, is independent and normally distributed with mean of 0 and variance τ_β. This variance allows for variation in the school-level mean mathematics scores.

To better understand the model formulated above, we write the combined model, which joins the three levels:

$$Y_{ijk} = \gamma_{000} + \gamma_{100}D_{2ijk} + \gamma_{200}D_{3ijk} + r_{0jk} + u_{00k} + \varepsilon_{ijk}. \qquad (10.4d)$$

From the CTT perspective, $\gamma_{000} + \gamma_{100}D_{2ijk} + \gamma_{200}D_{3ijk} + r_{0jk} + u_{00k}$ is the true score for subscale i for student j in school k, and ε_{ijk} is the random measurement error. The true score consists of three sources of variation: subscale, student, and school. Since two dummy variables, D_{2ijk} and D_{3ijk}, are indicators of subscale 2 and 3 and do not depend on student (j) and school (k), the fixed effects part of the true score, $\gamma_{000} + \gamma_{100}D_{2ijk} + \gamma_{200}D_{3ijk}$, can be interpreted as subscale difficulties, and the random effects, $r_{0jk} + u_{00k}$, can be interpreted as student overall mathematics proficiency, which is represented as the sum of student proficiency relative to the school mean and the school mean relative to the grand mean. Since the subscale difficulties differ and the errors are assumed to have homogeneous variances, the HLM model formulated above assumes an essentially parallel structure.

Results of fitting this three-level model to the example data are presented in Table 10.3. Note that by default the model is estimated using full maximum likelihood (FML) as opposed to restricted maximum likelihood (REML) (the default for the two-level HLM).[5]

The estimate of the student-level reliability, denoted as α_π, is obtained as .768, which is slightly lower than the estimate obtained from the conventional approach (α = .798), which ignored the nested data structure. The value .768 was computed by substituting the estimates of variances ($\hat{\tau}_\pi$ = 2.998 and

TABLE 10.3 Results from Three-Level Psychometric Model

a. Fixed Effects

	Estimate	Standard error	*t*-ratio	*df*	*p*-value
γ_{000}	5.547	.158	35.180	29	<.001
γ_{100}	−.731	.041	−18.057	9933	<.001
γ_{200}	−.050	.041	−1.230	9933	.219

b. Variance Components

	Estimate	Standard error	*df*	Chi-Square	*p*-value	Reliability
Level 1						
σ^2	2.717	.047				
Level 2						
τ_π	2.998	.098	3282	9358.894	<.001	.768
Level 3						
τ_β	.683	.188	29	527.171	<.001	.936

c. Model Summary

Deviance	# of parameter estimated
43053.238	6

$\hat{\sigma}^2$ = 2.717) and the number of subscales within student (n = 3) into the following formula:[6]

$$\hat{\alpha}_\pi = \frac{\hat{\tau}_\pi}{\hat{\tau}_\pi + \hat{\sigma}^2/n} = \frac{2.998}{2.998 + 2.717/3} = .768. \qquad (10.5)$$

This formula, where n is the number of test components, corresponds to the general "reliability" formula in multilevel modeling (e.g., Raudenbush & Bryk, 2002, p. 270). The reduction in the reliability coefficient occurred because the reliability obtained from a traditional model that ignores nesting is inflated due to the correlation of scores observed within the same school. In fact, the estimate of intraclass correlation (ICC), a correlation between the latent adjusted means within the same school is

$$\text{cor}(\pi_{0jk}, \pi_{0j'k}) = \frac{\tau_\beta}{\tau_\beta + \tau_\pi} = .186$$

for $j \neq j'$, which is not an ignorable magnitude of correlation (for more detailed discussion on the impact of the magnitude of the ICC, see, for example, Snijders & Bosker, 1999, p. 46).

The reliability of the school mean ($\bar{Y}_{\bullet\bullet k}$) was very high at .936. This value is computed by taking the average of the reliability coefficients for each school,

$$\alpha_\beta = \frac{1}{K}\sum \alpha_{\beta k}, \tag{10.6}$$

where

$$\alpha_{\beta k} = \frac{\hat{\tau}_\beta}{\hat{\tau}_\beta + \hat{\tau}_\pi / J_k + \hat{\sigma}^2 / J_k n}, \tag{10.7}$$

which were presented in Equations 10.14 and 10.16 in Raudenbush et al. (1991).[7] Here, J_k is the number of students for school k, and other terms have been defined above. In the formula, the school-level reliability, α_β, approaches one as J_k goes to infinity. In our data, the average number of students per school is 110; hence, the reliability of the school mean mathematics scores is very high at .936.

The implication of this difference in reliability at the school- versus student-level is that we can be much more confident in assigning a mathematics proficiency score to the school as a whole than any given student within the school. In other words, we more confidently can differentiate and rank order schools in terms of school performance on mathematic achievement than we can distinguish students within schools.

Measurement Models by Multivariate Three-Level Models

In this section, use of multivariate HLM for MMM is demonstrated. As mentioned earlier, the measurement model we considered in the previous section corresponds to the essentially parallel measures CTT structure, which requires that relatively strong assumptions be met (true scores are different only by test component difficulties, and equal component reliabilities). Not all of these assumptions are necessary for the scores to be interpretable; therefore, we may wish to relax some of the assumptions and test the fit empirically. We will use the multivariate module of HLM[8] software for this purpose. For other ways of applying multivariate HLM, such as to multiple outcomes and to repeated measures designs, the reader is referred to Raudenbush and Bryk (2002) and Thum (1997).

The first model we consider is the unrestricted model, where no structure is given to the variances/covariances of the level-one subtest true scores. This most unrestricted model imposes no structure on the covariance matrix and serves as the baseline model for comparison to more restrictive models.[9] The second model allows the level-one error variance to differ over the three subscales, while the covariances among the three

subscales are constrained to be the same. The weaker assumptions of this model correspond to the essentially tau-equivalent CTT model. Third, the homogeneous level-one error variance model is presented. The homogeneous level-one error variance model has only two covariance parameters: the level-one error variance and the level-two between-cluster variance. Thus, it is the simplest model and has the most restrictive covariance structure. The homogeneous level-one error variance model corresponds to an essentially parallel measure CTT model where error variances of measurement are homogeneous but true score means are heterogeneous. It also corresponds to the previous three-level psychometric model that was fit through a univariate approach.

Unrestricted Model

The level-one model consists of two equations. The first equation links the incomplete data vector (i.e., observed data) and the complete data vector (observed data and missing data) via a set of indicator variables so that

$$Y_{ijk} = IND1_{ijk}Y^*_{1jk} + IND2_{ijk}Y^*_{2jk} + IND3_{ijk}Y^*_{3jk} \tag{10.8a}$$

for $i = 1,\ldots, 3$, $j = 1,\ldots, 3312$, and $k = 1,\ldots, 30$, where Y_{ijk} is the ith observation of mathematics achievement subscale score for student j in school k, and through the three indicator variables (*IND1*, *IND2*, *IND3*) is linked to the complete data vector ($Y^*_{1jk}, Y^*_{2jk}, Y^*_{3jk}$) of the three subscale scores for student j in school k. *IND1* is the indicator variable for subscale 1 (measurement) of mathematics: it takes the value of one if the observation is from subscale 1 of mathematics and zero if not. Similarly, *IND2* and *IND3* are indicator variables for subscale 2 (algebraic thinking) and subscale 3 (data analysis/probability). Thus, the indicator variables differentiate which of the three subscales the observed value of Y_{ijk} measures.

The second part of the level-one model describes the measurement model for the outcome variables:

$$Y^*_{pjk} = \pi_{0jk} + \pi_{1jk}D2_{pjk} + \pi_{2jk}D3_{pjk} + e_{pjk} \quad (p = 1, 2, \text{and } 3), \tag{10.8b}$$

where Y^*_{pjk} is the pth subscale score student j in school k would have displayed if the score had been observed, π_{0jk} is the true score of subscale 1 for student j in school k, π_{1jk} is the difference score between subscales 1 and 2 for student j in school k, and π_{2jk} is the difference score between subscales 1 and 3 for student j in school k. $D2_{ij}$ is an indicator taking on a value of one if the response i belongs to subscale 2 (algebraic thinking) and zero otherwise, and $D3_{ij}$ is an indicator taking on a value of one if the response i belongs to subscale 3 (data analysis/probability) and 0 otherwise. The re-

sidual, e_{pjk}, is a random error associated with the pth subscale score, Y^*_{pjk}, an element of complete latent data. The second equation can be represented succinctly in matrix form as a multivariate linear model with 3 outcome variables:

$$\mathbf{Y}^*_{jk} = \begin{pmatrix} Y^*_{1jk} \\ Y^*_{2jk} \\ Y^*_{3jk} \end{pmatrix} = \begin{pmatrix} 1 & D2_{1jk} & D3_{1jk} \\ 1 & D2_{2jk} & D3_{2jk} \\ 1 & D2_{3jk} & D3_{3jk} \end{pmatrix} \begin{pmatrix} \pi_{0jk} \\ \pi_{1jk} \\ \pi_{2jk} \end{pmatrix} + \begin{pmatrix} e_{1jk} \\ e_{2jk} \\ e_{3jk} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{0jk} \\ \pi_{1jk} \\ \pi_{2jk} \end{pmatrix} + \mathbf{e}_{jk}. \quad (10.8c)$$

We assume that $\mathbf{e}_{jk}$ has a multivariate normal distribution with a mean of vector $\mathbf{0}$ and an arbitrary 3×3 covariance matrix,

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ \delta_{12} & \delta_{22} & \delta_{23} \\ \delta_{13} & \delta_{23} & \delta_{33} \end{pmatrix}, \quad (10.8d)$$

which involves six unique parameters. Then at level two, to let this covariance matrix, Δ, represent the overall between-student (within-school) variances and covariances among the three subtests, we constrain all the coefficients in Equation 10.8b (π's). That is,

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} \\ \pi_{1jk} &= \beta_{10k} \\ \pi_{2jk} &= \beta_{20k} \end{aligned}, \quad (10.8e)$$

The above model specification simply formulates the standard multivariate linear model for the complete data vector. We can see this by obtaining the combined model for the complete data vector, $\mathbf{Y}^*_{jk}$, by combining Equations 10.8c and 10.8e:

$$\mathbf{Y}^*_{jk} = A\beta_k + \mathbf{e}_{jk}, \quad (10.8f)$$

where

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

and $\beta_k = (\beta_{00k}, \beta_{10k}, \beta_{20k})^T$. From this Equation, we can confirm that the variance-covariance matrix, Δ, in fact, represents the dispersion of the complete data, $\mathbf{Y}^*_{jk}$, within schools, i.e., $\Delta = \text{var}(\mathbf{e}_{jk}) = \text{var}(\mathbf{Y}^*_{jk} \mid \beta_k)$. Note

that consistent with the absence of a specific measurement model, no structure has been imposed on this covariance matrix. Later models will differ in this regard.

The level-three units are schools. The level-three model that we formulate assumes only that the intercept varies across schools. This implies that some schools have higher average mathematics proficiency than others. Therefore,

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + u_{00k} \\ \beta_{10k} &= \gamma_{100} \\ \beta_{20k} &= \gamma_{200} \end{aligned}, \quad (10.8g)$$

where u_{00k} are independent, normally distributed with mean of 0 and variance of τ_β. Thus, the overall covariance structure for the complete data, $\mathbf{Y}^*_{jk}$, is

$$\text{var}(\mathbf{Y}^*_{jk}) = E[\text{var}(\mathbf{Y}^*_{jk} \mid \beta_k)] + \text{var}[E(\mathbf{Y}^*_{jk} \mid \beta_k)] = \Delta + A\,\text{var}(\beta_k)A^T = \Delta + \tau_\beta J_3,$$

where J_3 is a 3×3 matrix of all 1's.[10]

Results of fitting this model to the illustrative data are summarized in Table 10.4. The number of parameters estimated is 10 for this unrestricted model; three for fixed effects, six for the unique variances and covariances in Δ matrix, and τ_β, the variance of the first subscale (which is the reference group) at the school level.[11]

Heterogeneous Level-One Variance Model

The second model is the heterogeneous level-one variance model. For this model, the first equation at level one is the same as the unrestricted model (Equation 10.8a). The second equation of the level-one model includes the error term, e_{pjk},

$$Y^*_{pjk} = \pi_{0jk} + \pi_{1jk}D2_{pjk} + \pi_{2jk}D3_{pjk} + e_{pjk} \quad (p = 1,2,3), \quad (10.9a)$$

but here the e_{pjk} are independent and normally distributed with mean of 0 and constant variance of σ^2_1, σ^2_2, and σ^2_3 for the three subscales, respectively. The level-two equations are

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + r_{0jk} \\ \pi_{1jk} &= \beta_{10k} \\ \pi_{2jk} &= \beta_{20k} \end{aligned}, \quad (10.9b)$$

TABLE 10.4 Results for the Unrestricted Model for Three Subscales

a. Fixed Effects

	Estimate	Standard error	*t*-ratio	*df*	*p*-value
γ_{000}	5.548	.158	35.176	29	<.001
γ_{100}	−.731	.040	−18.191	3311	<.001
γ_{200}	−.050	.040	−1.240	3311	.215

b. Variance and Covariance Components

	Estimate	Standard error
Level 1 & 2		
var(e_{0jk})	5.684	.140
var(e_{1jk})	5.926	.110
var(e_{2jk})	5.535	.137
cov(e_{0jk}, e_{1jk})	3.128	.115
cov(e_{0jk}, e_{2jk})	2.935	.146
cov(e_{1jk}, e_{2jk})	2.932	.112
Level 3		
τ_β	.684	.188

c. Summary

Deviance	# of parameter estimated
43044.647	10

Note: The heading "Level 1 & 2" indicates that r_{jk} represents the student-level variability of a multivariate outcome, although it was used as the level-1 term in the model equation.

where r_{0jk} is normally distributed with mean of 0 and variance of τ_π. The level-three equations are exactly the same as for the unrestricted model (Equation 10.8g). Thus, the covariance matrix Δ is now structured as

$$\Delta = \begin{pmatrix} \tau_\pi + \sigma_1^2 & \tau_\pi & \tau_\pi \\ \tau_\pi & \tau_\pi + \sigma_2^2 & \tau_\pi \\ \tau_\pi & \tau_\pi & \tau_\pi + \sigma_3^2 \end{pmatrix}. \quad (10.9c)$$

The results for the heterogeneous level-one variance model for the sample data are presented in Table 10.5. Note that the number of estimated parameters in the heterogeneous model is now eight, two more than the homogeneous model.

Homogeneous Level-One Variance Model

The last multivariate multilevel model to be fitted is the homogeneous level-one variance model, which is equivalent to the standard three-level HLM

TABLE 10.5 Results for the Heterogeneous Level-One Variance Model

a. Fixed Effects

	Estimate	Standard error	*t*-ratio	*df*	*p*-value
γ_{000}	5.547	.158	35.176	29	<.001
γ_{100}	−.731	.041	−17.986	9933	<.001
γ_{200}	−.050	.040	−1.246	9933	.213

b. Variance and Covariance Components

	Estimate	Standard error
Level 1		
σ_1^2	2.619	.089
σ_2^2	2.857	.094
σ_3^2	2.675	.090
Level 2		
τ_π	2.997	.098
Level 3		
τ_β	.684	.188

c. Summary

Deviance	# of parameter estimated
43049.851	8

demonstrated previously. In this model the equal error variance assumption is further imposed for the previous model. Model equations are unchanged from the previous model. Thus, in the second equation of the level-one model, Equation 10.9a, the errors e_{pjk} are still independent and normally distributed with mean of 0. However, the variances are now assumed to be equal for the three subscales. Thus, the covariance matrix Δ is now structured as

$$\Delta = \begin{pmatrix} \tau_\pi + \sigma^2 & \tau_\pi & \tau_\pi \\ \tau_\pi & \tau_\pi + \sigma^2 & \tau_\pi \\ \tau_\pi & \tau_\pi & \tau_\pi + \sigma^2 \end{pmatrix}. \quad (10.10)$$

The results of the data analysis assuming a homogeneous level-one variance model are presented in Table 10.6.

Notice that the parameter estimates both for fixed and for random effects, as well as the value of the deviance at convergence, are the same as the ones obtained from the standard three-level HLM (see Table 10.3), which were demonstrated earlier in this chapter. These results confirm that the MMM based on the multivariate HLM with homogenous level-one variance is equivalent to the MMM based on the standard univariate three-level HLM.

TABLE 10.6 Results for the Homogeneous Level-One Variance Model

a. Fixed Effects

	Estimate	Standard error	*t*-ratio	*df*	*p*-value
γ_{000}	5.547	.158	35.180	29	<.001
γ_{100}	−.731	.041	−18.057	9933	<.001
γ_{200}	−.050	.040	−1.230	9933	.219

b. Variance and Covariance Components

	Estimate	Standard error
Level 1		
σ^2	2.717	.047
Level 2		
τ_π	2.998	.098
Level 3		
τ_β	.683	.188

c. Summary

Deviance	# of parameter estimated
43053.238	6

Finally, we compare the three nested models by deviance tests to evaluate which model is most appropriate for the data. The results in Table 10.7 indicate that the homogeneous level-one variance model is the best fitting model among the three for these data, since the change of the deviance was not significant for either the heterogeneous level-one variance model or the unrestricted model at .05 level of significance (see Chapter 7 of this volume (McCoach & Black, 2008) for details on model comparison).

HGLM Approach for Multilevel Measurement Model

The previous section described multilevel measurement models for continuous measurement indicators. Although this assumption is suitable in certain circumstances, such as the illustrative analysis in previous sections, it is often the case that measurement indicators are categorical variables. In fact, the item-level response data in the illustrative data example consists entirely of dichotomous items. Therefore, if one wishes to use the items as measurement indicators, it will be more appropriate to treat measurement indicators as categorical variables. Other examples include ordered categorical items, such as Likert-type scale items in an attitude survey. In these cases, generalized multilevel linear or nonlinear models that are equivalent to well-known

TABLE 10.7 Summary of the Three Models by Considering Nested Data Structure

a. Deviance

Model	Parameters estimated	Deviance
1. Unrestricted model	10	43044.647
2. Heterogeneous level-1 variance model	8	43049.851
3. Homogeneous level-1 variance model	6	43053.238

b. Deviance Test

	Chi-square	*df*	*p*-value
Model 3 vs. Model 2	3.387	2	.182
Model 2 vs. Model 1	5.204	2	.072
Model 3 vs. Model 1	8.590	4	.071

item response models, such as the Rasch model or the two-parameter item response theory (IRT) model, can be specified for the data.

It is still an option to use a linear multilevel model for categorically-measured indicator variables. Specifically, for dichotomous indicator variables, a linear probability model can be fit using the untransformed 0, 1 dichotomous responses as the outcomes (item responses); thus, the predicted dependent variable corresponds to the probability of observing the event, such as supplying a correct answer rather than an incorrect answer. However, this is not a desirable approach for several reasons, such as possibility of out-of range predicted probability values. Chapter 6 of this volume (O'Connell, Goldstein, Rogers, & Peng, 2008) and Long (1997), for example, provide detailed discussions of the problems associated with the linear probability model.

A more desirable approach to modeling categorical indicator variables is to use a generalized linear model (GLM) extension of the multilevel linear model, such as the hierarchical generalized linear model (HGLM) (O'Connell et al., 2008; Raudenbush & Bryk, 2002), specifically, one using the logit link. In this section, we assume that the responses are scored dichotomously in a manner similar to our example data.[12] Let $Y_{ijk} = 1$ if the ith response is correct for student j of school k and $Y_{ijk} = 0$ otherwise, and let μ_{ijk} be the probability of $Y_{ijk} = 1$. This probability varies randomly across students. However, conditioning on this probability, we have $Y_{ijk}|\mu_{ijk} \sim$ Bernoulli with $E(Y_{ijk}|\mu_{ijk}) = \mu_{ijk}$ and $\text{var}(Y_{ijk}|\mu_{ijk}) = \mu_{ijk}(1 - \mu_{ijk})$. Then, a multilevel measurement model[13] can be written for the example data set as

$$\text{logit}(\mu_{ijk}) = \gamma_i + r_{jk} + u_k, \tag{10.11}$$

where γ_i is the effect of item i. Student ability variation within school, $r_{jk} \sim N(0, \sigma_r^2)$ and the school mean ability variation $u_k \sim N(0, \sigma_u^2)$ are assumed. This model is equivalent to the Rasch model, where $-\gamma_i$ is the item difficulty for item i and $r_{jk} + u_k$ is the trait level for person j in school k. The multilevel measurement model takes into account within-school (or between-students for each school) variability, as well as between-school variability, while the Rasch model is a single-level model that only considers between-students variability for all schools combined. This distinction is analogous to the difference between the three-level MMM discussed in the previous section relative to conventional CTT models. In fact, Equation 10.11 can be simplified to a two-level model by not considering the level-three variation,

$$\text{logit}(\mu_{ij}) = \pi_i + r_j. \tag{10.12}$$

In this case, the model is equivalent to the Rasch model, where $-\pi_i$ is the item difficulty for item i and r_j is the trait level for person j.[14]

Models in Equations 10.11 and 10.12 are HGLMs based on a logit link function, with a three-level HGLM for Equation 10.11 and a two-level HGLM for Equation 10.12. The quantity being predicted is the log of the odds of getting item i correct for the jth child in the kth school; the model in Equation 10.12 assumes there are no school differences. However, some constraints need to be imposed to identify the parameters of the model. Several different ways to parameterize the model have been suggested, as well as different estimation methods and optimization methods. For example, Kamata (2001) demonstrated that this can be modeled in the framework of HGLM using one item as a reference item and including an intercept term. Also, it is possible to estimate parameters in the model by constraining the mean item difficulties to be zero without specifying a reference item (e.g., Cheong & Raudenbush, 2000).

Parameter estimation can be accomplished by delete several different methods, including the penalized quasi likelihood (PQL), Laplace approximation, Gaussian numerical integration of log-likelihood, and fully Bayesian Markov Chain Monte Carlo (MCMC) methods. Rijmen, Tuerlinckx, Meulders, Smits, and Balazs (2005) compared these estimation methods and demonstrated that all estimators performed equally well in reasonable conditions. In the examples below, we use the HLMv6 software, which utilizes PQL estimation. Roberts and Herrington (2005) demonstrated how to set up and analyze data for these models in several different software packages.

Equation 10.11 is fit by a three-level HGLM for the example data, which is analogous to the homogeneous level-one variance model in the linear case, except that the outcomes (items) are treated as dichotomous rather

than continuous. In the illustrative data, there are 22 items. Therefore, the level-one equation is the item-level model and can be written as

$$\text{logit}(\mu_{ijk}) = \pi_{0jk} + \sum_{q=1}^{21} \pi_{qjk} D_{qijk}, \tag{10.13}$$

where D_{qijk} is the qth indicator variable that takes a value of one if $q = i$ for item i. There is no error term in Equation 10.13 because it is absorbed by the link function (O'Connell et al., 2008). One item is used as a reference item, and the other item difficulties are assessed relative to the reference item. Thus, $q = 1, \ldots, 21$, rather than up to 22 (8 + 7 + 7 = 22 items). Then, the level-two equations are

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + r_{0jk} \\ \pi_{qjk} &= \beta_{q0k} \end{aligned}, \tag{10.14}$$

where $q = 1, \ldots, 21$. The slopes are not random because the item difficulties are assumed to be equal across individual students. The level-three equations are

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + \mu_{00k} \\ \beta_{q0k} &= \gamma_{q00} \end{aligned}. \tag{10.15}$$

The slopes are not random because the item difficulties are assumed to be equal across schools. As a result, γ_{000} is the difficulty of the reference item, and γ_{q00} is the difference between item i (for $q = i$) and the reference item in their difficulties. The ability of student j in school k is $r_{0jk} + u_{00k}$. Both r_{0jk} and u_{00k} are assumed to be normally distributed with means of zero and unknown variances.[15] The results of this model are presented in Table 10.8.

The difficulty of item 40 (the reference item) was estimated to be .344 (γ_{000}) logits, indicating it is .344 logits higher than the mean ability. Other values indicated in Table 10.8a are differences in their difficulties compared to item 40. For example, item 6 is more difficult than item 40 by .178; thus, its difficulty is .344 + .178 = .522. On the other hand, item 7 is easier than item 40 by .408, so its difficulty is .344 − .403 = −.059. Notice that $\text{se}(\gamma_{i00})$ for item 40 is the standard error for the item's difficulty while the standard errors for the remaining estimated parameters are standard errors for the difference in difficulty from that of item 40. For this model, the variances of the student abilities are provided in the bottom panel of Table 10.8b. The within-school variance was estimated as $\sigma^2 = \text{var}(r_{0jk}) = .694$, while the between-school variance was estimated as $\tau = \text{var}(u_{00k}) = .162$. This implies that the intraclass correlation in latent mathematics ability is

$$\frac{\tau}{\sigma^2 + \tau} = .162 / (.694 + .162) = .189.$$

In other words, 18.9% of the variability in mathematics ability can be ascribed to differences between schools as opposed to variability among students within schools.

TABLE 10.8 Results of Example Data Analysis by HGLM – Unconditional Model

a. Fixed Effects

Item	γ_{i00}	se(γ_{i00})
Measurement		
6	.178	.054
7	−.403	.053
8	.023	.053
10	−.028	.053
28	.061	.053
31	−1.342	.055
39	.883	.057
40*	.344	.084
Algebraic thinking		
14	−.555	.053
16	−.265	.053
17	−.785	.053
19	−1.266	.055
21	−.006	.053
35	−.508	.053
36	.399	.054
Data analysis		
4	.377	.054
25	−.485	.053
26	.507	.055
29	−.977	.054
30	−.201	.053
33	−.874	.054
38	1.061	.058

b. Random Effects

	Estimate	Standard Error
Level 2		
var(r_{0jk})	.694	.016
Level 3		
var(u_{00k})	.162	.047

* This item was used as the reference item in the model. Therefore, the parameter listed for this item is the estimate of the intercept γ_{000}.

Model Extensions with a Covariate

Both HLM and HGLM models can be extended easily to include covariates and additional variance and covariance components. For example, assume we have an additional person-level predictor (X_{jk}) in the three-level HGLM measurement model and that our interest is in the main effects of the additional person-level predictor and the interaction effect between the additional predictor and the item indicator ($D_{ijk}X_{jk}$). We can still use Equation 10.13 as the level-one model. The level-two equations become

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + \beta_{01k}X_{jk} + r_{0jk} \\ \pi_{qjk} &= \beta_{q0k} + \beta_{q1k}X_{jk} \end{aligned}, \qquad (10.16)$$

where $q = 1,\ldots, 21$. Here, β_{01k} is the main effect of the person-level predictor, and β_{q1k} is the person by item interaction effect. Furthermore, if we also are interested in the random variation of the person by item interaction effect across the level-three units, the level-three equations become

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + u_{00k} \\ \beta_{01k} &= \gamma_{010} \\ \beta_{q0k} &= \gamma_{q00} \\ \beta_{q1k} &= \gamma_{q10} + u_{q1k} \end{aligned}, \qquad (10.17)$$

where $q = 1,\ldots, 21$ and u_{q1k} is the random effect of the interaction effect. In addition to the fixed effects (γ_{000}, γ_{010}, γ_{q00}, and γ_{q10}), the variance and covariance components, var(r_{0jk}), var(u_{00k}), var(u_{q1k}), and cov(u_{00k}, u_{q1k}), are estimated.

If X_{jk} is a dichotomous variable that represents two subpopulations of test examinees, the interaction effect γ_{q10} is a differential item functioning (DIF) parameter that enables one to detect potential item bias. When subgroups of examinees have different probability of answering an item correctly, given the same level of abilities, we say the item displays DIF.[16] Furthermore, this three-level formulation of the model is equivalent to the "random effect DIF model" presented by Cheong (2006) and Kamata, Chaimongkol, Genc, and Bilir (2005). In this context, one's interest is in estimating the magnitude of γ_{q10} (the mean magnitude of DIF across schools) and var(u_{q1k}) (the randomly varying DIF magnitude across schools). Also, cov(u_{00k}, u_{q1k}) indicates how the mean performance of students and DIF magnitude are related at the school level.

For demonstration, one student-level variable, enrollment in a free or subsidized lunch program, is used. Fifty-seven percent of the students in

the sample were enrolled in free or subsidized lunch programs. Using the HGLM option in the HLM software, we fit the model by arbitrarily treating the last item (item 40) in the measurement subscale as a reference item. We needed further constraints to fix the magnitude of the fixed interaction effect (DIF magnitude; γ_{i10} in Equation 10.17) for identification reasons. Our preliminary data exploration indicated the magnitude of the DIF for the third item in the measurement subscale was near zero ($\gamma_{i10} = .0001$ for this item). Therefore, its effect was constrained to be zero, and the parameter was dropped from the model. Accordingly, $\text{var}(u_{i1k})$ was also constrained to be zero for this item.

Seven items displayed statistically significant DIF (DIF estimates were larger than twice their standard errors). These items are indicated by asterisks in Table 10.9. Each had negative values, indicating that students who participated in free or subsidized lunch programs had significantly lower odds of correct answers for the indicated items, given the same level of ability. For item 33 in the data analysis subscale, for example, the odds of a correct answer for students who received free or subsidized lunch was only 70% (exp[−.351] = .704) as high as the odds of success for students who were not receiving free or subsidized lunch. For the seven items that displayed significant DIF, random effects were further estimated in a separate model.[17] Six of these items displayed statistically significant variability, indicating variation in the degree of DIF across schools. For example, the estimate of $\text{var}(u_{i1k})$ was .230 for item 33. In conjunction with the estimate of the fixed effect, it can be interpreted that the 95% of logits for DIF on item 33 are in the range of $-.351 \pm 1.96\sqrt{.230} = [-1.291, .589]$, assuming the normality of the distribution of DIF across schools.

By examining the estimated $\text{cor}(u_{00k}, u_{i1k})$ in Table 10.9, we see that the correlations are all positive among the seven items with significant DIF. Positive correlations indicate that DIF was higher for schools with higher mean performance. These seven items had negative DIF magnitudes, indicating students with free or subsidized lunch had lower odds of correct answer, given the same level of ability. At a glance this may seem counter-intuitive. However, given that the mean interaction effect (DIF) was a negative value, stronger interaction effects are actually values of DIF closer to zero. In other words, the interaction effect resulted in DIF values that were more positive (or closer to zero) in schools with higher mean performance.

Summary and Limitations of the HLM/HGLM Approach

To this point, we have demonstrated how to fit measurement models using hierarchical linear and generalized linear models. These approaches are different in how they treat measurement indicator variables; consequently, their link functions are different. However, they are very similar in other ways. Therefore, HLM and HGLM approaches are discussed together in this section.

The basic approach of these multilevel measurement models is to conceive of the measured variables or item responses as the lowest level of a three-level model with levels corresponding to the observed measures (level one), persons (level two), and groups (level three). In the case where there are no groups (i.e., ignoring school membership, or when school differences were not part of the research design), the model reduces to a simple random intercept model, where the random intercept now accounts

TABLE 10.9 Estimates of Fixed and Random Effects for the Interaction Effect

	Fixed effects		Random effects	
Item	γ_{i10}	se	$\text{var}(u_{i1k})$	$\text{cor}(u_{00k}, u_{i1k})$
Measurement				
6	−.247*	.113	.072**	.134
7	.000**	—		
8	.024	.133		
10	−.297*	.110	.059**	.436
28	−.149	.111		
31	−.066	.107		
39	−.307*	.117	.078**	.573
40	.000**	—		
Algebraic thinking				
14	.139	.128		
16	−.155	.103		
17	−.284*	.117	.144**	.453
19	−.069	.108		
21	.007	.110		
35	−.072	.105		
36	−.417*	.105	.010	.476
Data analysis				
4	−.199	.113		
25	−.182	.107		
26	−.264*	.114	.091**	.643
29	−.040	.122		
30	−.037	.142		
33	−.351*	.128	.230**	.499
38	.111	.124		

* Magnitudes are greater than twice the standard errors.
** Magnitudes are significant at $\alpha = .05$ based on chi-square test.

for the dependence among the repeated measures (items) within person and, hence, constitutes the latent factor or trait. In the linear case (with continuous measurement indicators), this corresponds to a confirmatory factor model assuming equal factor loadings, or an essentially tau-equivalent structure that assumes equal discrimination. For dichotomous items, a random intercept model produces the well-known Rasch model, which also assumes equal discrimination across items. In both cases, differences in the difficulties of test components are modeled by including the fixed effects of dummy variables for each measure.

One advantage of formulating these models in the HLM/HGLM framework as compared to the traditional (single-level) CTT or IRT frameworks is that these measurement models can be extended to allow for variance components in the latent factors or traits at both the individual and group level. When the full three-level model is specified, this allows for variability in the factor means (item subscales) across groups, as well as within-group variability in individual levels on the latent factor. As such, the total variance of the latent factor can be decomposed into between- and within-school components. The intraclass correlation obtained from this decomposition of the latent variable typically exceeds the magnitude of the intraclass correlations for the measured variables, reflecting disattenuation for measurement error (Raudenbush et al., 1991). Another advantage is that observed predictors can be included in the model at either the individual and/or group level to explain the two components of variance.

There are, however, some limitations to incorporating measurement models into HLM or HGLM. One limitation is the assumption that the discrimination power (factor loadings) are equal (or at least known a priori) for all test components. Ideally, these model parameters could be estimated directly from the data, just as typically is done in confirmatory factor analysis and two-parameter item response models. Many empirical applications of linear factor models and item response models where the factor loadings differ across the observed measures suggest the need for this modeling flexibility.

By assuming equal factor loadings, we are assuming that the relationships between the observed measures and the latent factor are equivalent across all test components. In test-scoring, it is considered desirable for a measurement instrument to possess such a property (see e.g., Embretson & Reise, 2000). Thus, when this equivalence assumption is consistent with the data, it provides for a parsimonious and useful measurement model for the instrument. On the other hand, a less restrictive model would allow the factor loadings or item discriminations to vary, rather than constraining them to be equal a priori. This constraint is more difficult to impose for binary items because a model that contains item-specific discrimination parameters is not a hierarchical generalized linear model anymore (Rijmen et al., 2003). More generally, a fundamental limitation of the HLM/HGLM approach is that the random coefficients are related to the observed repeated measures via a design matrix, which, by definition, must consist of known values (see Bauer, 2003). The random intercept that constitutes the latent factor or trait is defined by inserting a column of ones into the design matrix for the random effects. To overcome this limitation, one must leave the HLM/HGLM framework so that the design matrix can be replaced by a matrix that allows the inclusion of both known values (e.g., covariates) and unknown values (e.g., factor loadings or discrimination parameters).

For binary items, Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003) and Rijmen and Briggs (2004) provide an example of such an approach using a non-linear mixed model. According to their approach, a 2PL IRT model can be modeled by treating a logit of the probability of the response as a linear function. We assume that the distribution of the latent trait is an arbitrary distribution, such as a standard normal distribution. Also, we assume that the probability of observing 1 rather than 0 for the dependent variable is defined by the cumulative standard logistic distribution. One limitation is that the available software (e.g., PROC NLMIXED in SAS) is limited to the formulation of two-level models. (This does not preclude the inclusion of level-three covariates. In fact, we do not have to distinguish the level of hierarchy for fixed effects, such as covariates in the model.) Thus, we cannot estimate the variance and covariance components of the level-three model, such as $\text{var}(u_{00k})$, $\text{var}(u_{q1k})$, and $\text{cov}(u_{00k}, u_{q1k})$ from Equation 10.17.

An additional limitation of the HLM/HGLM approach concerns the simultaneous modeling of several latent variables. Multiple latent variables can, in fact, be estimated by removing the intercept from the model and estimating random effects for predictors coded one or zero to differentiate groupings among the observed measures or items (see e.g., Cheong & Raudenbush, 2000; Kamata & Cheong, 2008; Raudenbush et al., 1991). However, the structure applied to the covariance matrix among these latent variables often is quite limited. Typically, the covariances would be left unstructured, indicating that each latent factor is correlated with every other latent factor and that there are no structural relations between them. The need to allow for such effects is demonstrated by the popularity of structural equation models that include regressions among latent variables. Both predictors and outcomes can be defined as latent variables and estimates of the effects can be obtained that are unbiased by measurement error.

Another limitation of the HLM/HGLM approach to incorporating latent variables in hierarchical models is that it imposes a highly structured model on the within-group and between-group variability of the observed measures. Specifically, the HLM/HGLM approach assumes that the group means of these measures vary randomly across groups but that the variation in the factor means entirely accounts for this variability. As will be discussed

later in more detail, this assumption implies that (1) the same factor structure (e.g., dimensionality) holds both within-groups and between-groups; (2) the factor loadings (or discrimination parameters) are identical at both levels of the model; and (3) there are no group-mean differences in the observed measures that are not due to variation in the factor means (or other covariates included in the model). Although HLM/HGLM approach is a parsimonious and easily understood model, these restrictions may not always be consistent with theory or hold in practice. Ideally, these restrictions should be tested and relaxed when inconsistent with the data.

Given these limitations of the HLM/HGLM approach to the design and analysis of measurement models, we will conclude this section by discussing some additional alternative methods. One alternative is the Generalized, Linear, Latent and Mixed Model (GLLAMM) of Skrondal and Rabe-Hesketh (2004), an add-on program for STATA. This model allows for the estimation of factor loadings or discrimination parameters, the specification of structural relations between latent variables, and differences in the between-group and within-group model structure. This model is very general. However, because the estimation requires numerical integration, specifications including several latent variables and/or other random effects can be computationally intensive. In fact, GLLAMM allows us to formulate the same model used in the example data analyses based on Equations 10.16 and 10.17, along with discrimination parameters. In this example, however, we had eight random effects at level three of the model (seven random DIF and one school-level variance of latent abilities), and this number of random effects, unfortunately, makes numerical integration computationally impractical. One strategy to avoid such a large number of random effects is to simplify the model, for instance, by estimating a random DIF effect for one item at a time, which results in two random effects at level three but seven separate data analyses. This approach is much more feasible computationally; however, all random DIF effects except the one being investigated are constrained to be zero in each analysis. This may or may not be a reasonable assumption. It is hoped that improved computational algorithms for these types of model will be available in the future.

Another approach that has the same flexibility, with the exception that it is applicable only for linear models, is the two-level structural equation model (SEM). It is this approach that we discuss in greater detail in the next section, relaxing the assumptions of the HLM/HGLM approach discussed previously.

TWO-LEVEL STRUCTURAL EQUATION MODEL

SEM is a multivariate method that generalizes regression, path analysis, and factor analysis. At its core, SEM represents the integration of measurement

models (e.g., factor analysis) with simultaneous equations (e.g., path analysis). It allows for the definition of multiple latent variables and structural relations among the latent variables. Many HLMs can be fit as single-level SEMs using a multivariate approach wherein the level-one observations are construed to be separate variables. This relation is well-known for growth models (Willett & Sayer, 1994) but also holds more generally (Bauer, 2003; Curran, 2003). As an alternative to the single-level approach, the covariance matrix of the measured variables can be modeled simultaneously at both the within- and between-group level in a two-level SEM (Goldstein & McDonald, 1988; McDonald & Goldstein, 1989; Muthén, 1994; Muthén & Satorra, 1995). To demonstrate this extension, we first describe the single-level (standard) SEM and then proceed to allow for an additional level of nesting, such as data obtained from children within classrooms or children within schools.

The measurement model of the linear SEM can be defined by the following equation:

$$\mathbf{y}_j = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j. \tag{10.18}$$

This is effectively just a linear regression of the vector of n observed variables, $\mathbf{y}_j$, on the latent variables, $\boldsymbol{\eta}_j$ for person j. $\mathbf{y}_j$ is an $n \times 1$ vector that contains scores or responses to i measurement indicators, while $\boldsymbol{\eta}_j$ is a $P \times 1$ vector that contains latent scores for P latent factors. As such, the notation reads as follows: $\boldsymbol{\nu}$ are intercepts ($n \times 1$ vector), $\boldsymbol{\Lambda}$ are slopes (factor loadings) ($n \times P$ matrix), and $\boldsymbol{\varepsilon}_j$ are residuals ($n \times 1$ vector). For the previous illustrative data analysis with continuous measurement indicators (HLM formulation), $n = 3$ with 3 subscales and $P = 1$ with one latent factor (mathematics ability) to be measured. Of importance, the residuals, $\boldsymbol{\varepsilon}_j$, are assumed to be normally distributed with means of zero and $n \times n$ covariance matrix $\boldsymbol{\Theta}$ (often, but not necessarily, assumed to be diagonal, reflecting independent residuals or local independence).

The latent variable model of the SEM can then be written as

$$\boldsymbol{\eta}_j = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_j + \boldsymbol{\zeta}_j. \tag{10.19}$$

This, too, is simply a linear regression; only this time, latent variables are regressed on other latent variables. The intercepts and slopes of this latent variable regression are given by $\boldsymbol{\alpha}$ ($P \times 1$ vector) and $\mathbf{B}$ ($P \times P$ matrix), respectively, and $\boldsymbol{\zeta}_j$ are the residuals ($P \times 1$ vector). The residuals are assumed to be normally distributed with means of zero and $P \times P$ covariance matrix, $\boldsymbol{\Psi}$. These latent variable residuals also are assumed to be uncorrelated with the residuals from the measurement model. Note that it is not always necessary to include latent variable regressions in an SEM. If there is no latent

variable regression in an SEM (for instance, in a confirmatory factor analysis), the **B** matrix is a null matrix. Then, the intercepts, α, are interpreted simply as factor means, and ψ is the covariance matrix of the latent factors. However, including latent variable regressions (via a non-null **B** matrix) allows the researcher to examine structural relations that are unbiased by measurement error, including causal chains involving several latent variables (e.g., indirect or mediated effects).

These equations and assumptions imply that the mean vector and covariance matrix of the measured variables are

$$\mu = \nu + \Lambda(\mathbf{I} - \mathbf{B})^{-1}\alpha,$$

$$\Sigma = \Lambda(\mathbf{I} - \mathbf{B})^{-1}\psi\left[(\mathbf{I} - \mathbf{B})^{-1}\right]'\Lambda' + \Theta. \tag{10.20}$$

Here, **I** is a $P \times P$ identity matrix. If no latent variable regressions are present in the model (**B** is a null matrix), then these equations simplify considerably to

$$\mu = \nu + \Lambda\alpha,$$

$$\Sigma = \Lambda\psi\Lambda' + \Theta. \tag{10.21}$$

Readers may be familiar with these equations as giving the structure of a confirmatory factor model (see e.g., Bollen, 1989). To ease the exposition of the two-level SEM, it is the simpler model in Equation 10.21 that we will focus on here, though extensions to the full model in Equation 10.20 are straightforward.

In practice, maximum likelihood typically is used to find the estimates for the model parameters that most likely would have given rise to the observed data. In estimating the (single-level) SEM model, the log-likelihood is summed over individuals in the sample, requiring the assumption that the data vectors (responses to measurement indicators) for any two individuals are independent. Further detail on the single-level SEM may be sought from a number of excellent texts, including Bollen (1989), Kaplan (2000), or Kline (2005).

The two-level SEM differs from the foregoing single-level model in assuming that data are obtained from multiple individuals randomly sampled from each of many groups in the population. To account for the correlations among individuals within groups, it is assumed that the intercepts of the measured variables vary randomly over groups. The factor model can then be written as

$$\mathbf{y}_{jk} = \nu_k + \Lambda_W\eta_{jk} + \varepsilon_{jk}, \tag{10.22}$$

where k indexes group, and the subscripting of the intercept vector indicates that intercepts vary randomly over groups. Within groups, the latent factors are assumed to be normally distributed with mean vector α and covariance matrix ψ_W, and the residuals are assumed to be normally distributed with means of zero and covariance matrix Θ_W. A key assumption is that these covariance matrices are homogeneous across all groups. As such, for any given group k (i.e., fixing ν_k to a specific value), the (pooled) within-group covariance matrix is given by essentially the same equation as the standard SEM Equation 10.21, namely

$$\Sigma_W = \Lambda_W\psi_W\Lambda_W' + \Theta_W, \tag{10.23}$$

where the W subscript indicates "within-groups."

The key difference between the two-level SEM and the standard single-level SEM involves the additional component of variability due to the random intercepts. These intercepts are assumed to be independent of the other terms in Equation 10.22 and normally distributed across groups:

$$\nu_k \sim N(\nu, \Sigma_B). \tag{10.24}$$

Here, ν captures the average intercepts of the indicators over groups, and the covariance matrix Σ_B refers to the between-groups covariance, or the covariance due to group mean (intercept) differences. The insight behind multilevel SEM is to impose an additional factor structure on this covariance matrix (Ansari, Jedidi, & Dube, 2002; Goldstein & McDonald, 1988; McDonald & Goldstein, 1989; Muthén, 1994; Muthén & Satorra, 1995). The resulting equation is

$$\Sigma_B = \Lambda_B\psi_B\Lambda_B' + \Theta_B. \tag{10.25}$$

Similar expressions could be given for the full multilevel SEM with latent variable regressions. Note that while the structure applied to the within- and between-groups covariance matrices appears very similar, the differential subscripting of the matrices by W or B, respectively, indicates that the parameter estimates or even the factor structure of the model can differ between the two parts of the model.

To summarize, under the assumption that the groups differ only in their intercepts, the total covariance matrix can be partitioned into a (pooled) within-groups component, reflecting associations observed within groups, and a between-groups component, reflecting associations observed between groups (due to the group mean differences). A model is fit simultaneously to these two covariance matrices. Sample estimates of Σ_W and Σ_B can be computed and provide sufficient statistics for estimating the model by an

approximate maximum likelihood estimator (Muthén's maximum likelihood, or MUML; Muthén, 1995; Muthén & Satorra, 1995). More recently, a true full information maximum likelihood estimator also has become available in several SEM software programs.[18]

Relationship to the HLM Approach

Although superficially quite different, the two-level SEM and the HLM approach to fitting measurement models presented earlier in this chapter are, in fact, similar in many ways. Namely, both the two-level SEM and the HLM approach assume that the means of the observed variables vary randomly over groups. In the two-level SEM, it is the vector of intercepts that varies, whereas in the HLM approach, it is the vector of factor means. For example, the intercepts for every item can vary randomly in the SEM, but in HLM, only the factor mean scores can vary. The latter model is, in fact, more restrictive. This can be seen in equation form where the two-level SEM assumes that the group means for the measured variables, $\boldsymbol{\mu}_k$, can be expressed as

$$\boldsymbol{\mu}_k = \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_W \boldsymbol{\alpha}. \tag{10.26}$$

This follows from the fact that in Equation 10.22 only the intercepts vary over groups (and the latent factors are distributed identically over groups, with mean vector $\boldsymbol{\alpha}$). By adding and subtracting the mean vector for $\boldsymbol{\nu}_k$, Equation 10.26 can then be rewritten as

$$\boldsymbol{\mu}_k = \boldsymbol{\nu} + \boldsymbol{\Lambda}_W \boldsymbol{\alpha} + (\boldsymbol{\nu}_k - \boldsymbol{\nu}), \tag{10.27}$$

where the covariance matrix for the last term, the random component, is $\boldsymbol{\Sigma}_B$ and has the structure given in Equation 10.25. The HLM approach assumes that this random component actually arises due to group differences in the factor means. Thus, in the HLM approach, the model for the group means can be written as

$$\boldsymbol{\mu}_k = \boldsymbol{\nu} + \boldsymbol{\Lambda}_W \boldsymbol{\alpha} + \boldsymbol{\Lambda}_W (\boldsymbol{\alpha}_k - \boldsymbol{\alpha}), \tag{10.28}$$

where $\boldsymbol{\alpha}_k$ is the vector of randomly varying factor means. Notice that this model differs from Equation 10.27 only in that the term $(\boldsymbol{\nu}_k - \boldsymbol{\nu})$, reflecting random intercepts, has been replaced by the term $\boldsymbol{\Lambda}_W(\boldsymbol{\alpha}_k - \boldsymbol{\alpha})$, reflecting random factor means.

Since there are fewer latent factors than observed measures, the HLM approach reduces the number of random effects and results in a very parsimonious model. From Equation 10.28, we can derive the implied covariance matrix of $\boldsymbol{\mu}_k$ in the HLM approach to be

$$\boldsymbol{\Sigma}_B = \boldsymbol{\Lambda}_W \boldsymbol{\Psi}_B \boldsymbol{\Lambda}_W', \tag{10.29}$$

where $\boldsymbol{\Psi}_B$ now is interpreted as the covariance matrix for the random factor means. By comparing Equations 10.25 and 10.29, we can see that the HLM approach assumes that the factor loading matrix (and hence factor structure) is equal at the within and between levels and that there is no residual variability at the between level (i.e., $\boldsymbol{\Lambda}_B = \boldsymbol{\Lambda}_W$ and $\boldsymbol{\Theta}_B = \mathbf{0}$; Rabe-Hesketh, Skrondal, & Pickles, 2004). Thus, the HLM approach assumes a more restricted model than the SEM approach. Within the two-level SEM, one can fit a model with these same restrictions and test (via likelihood ratio) whether the restrictions are tenable for the data.

Two other advantages of the two-level SEM also are worth noting. One important advantage is that the factor loadings at both levels of the model can be estimated and need not be pre-specified by the analyst. This permits the estimation of two-level congeneric measurement models and measurement models including cross loadings, etc. Another advantage of the two-level SEM is that we can extend the preceding equations to allow for causal relations among latent factors at both the individual and group levels (similar to the single-level SEM). In contrast, the HLM approach typically assumes the covariance matrix among the latent factors to be unrestricted. We now demonstrate the features of the two-level SEM with our empirical example.

Example Analyses

To demonstrate some of the advantages of the two-level SEM, we now reanalyze the math scale data introduced earlier, comprised of the three subscale scores for mathematics proficiency. We used *Mplus* to fit these models (Muthén & Muthén, 2004), although other SEM software, such as LISREL and EQS, is equally capable of fitting multilevel SEMs. The three-level HLM analysis of the data reported earlier, with equal residual variances across items represented by Equations 10.4a–10.4c, can be written as a two-level SEM with the following specifications

$$\boldsymbol{\nu} = \begin{bmatrix} 0 \\ \gamma_{100} \\ \gamma_{200} \end{bmatrix}; \ \boldsymbol{\Lambda}_W = \boldsymbol{\Lambda}_B = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}; \tag{10.30}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \gamma_{000} \end{bmatrix}; \ \boldsymbol{\Theta}_W = \sigma^2 \mathbf{I}; \ \boldsymbol{\Theta}_B = \mathbf{0}; \ \boldsymbol{\Psi}_W = \begin{bmatrix} \tau_\pi \end{bmatrix}; \ \boldsymbol{\Psi}_B = \begin{bmatrix} \tau_\beta \end{bmatrix}.$$

Note that the symbols within these matrices correspond to the HLM notation used to report the earlier results and not the typical SEM notation. The fit of this model, as judged by the unrestricted model in HLM represented

by Equations 10.8a–10.8f, was seen to be good. However, it should be noted that the unrestricted model in a two-level SEM differs from the unrestricted model in HLM. The unrestricted HLM estimated earlier included the three means and six unique variances/covariances in Σ_W plus one school-level variance, totaling 10 parameters. From the standpoint of the two-level SEM, this is not really an unrestricted model, as it assumes that all six unique elements in Σ_B can be explained by the single parameter τ_β. That is, the mean and covariance structure for the three subscales implied by the unrestricted HLM (see notation defined for Equations 10.8d through 10.8f) are

$$\mu = \begin{bmatrix} \gamma_{000} \\ \gamma_{000} + \gamma_{100} \\ \gamma_{000} + \gamma_{200} \end{bmatrix}, \quad \Sigma_w = \Delta = \begin{bmatrix} \delta_{11} & & \\ \delta_{21} & \delta_{22} & \\ \delta_{31} & \delta_{32} & \delta_{33} \end{bmatrix},$$
$$\Sigma_B = \tau_\beta \mathbf{J}_3 = \begin{bmatrix} \tau_\beta & & \\ \tau_\beta & \tau_\beta & \\ \tau_\beta & \tau_\beta & \tau_\beta \end{bmatrix}. \tag{10.31}$$

Thus, this "unrestricted" model actually imposes a highly restrictive structure on the between-groups covariance matrix. In contrast, in the unrestricted two-level SEM, Σ_W and Σ_B both are estimated freely (each including six unique and unconstrained elements), along with the means for the three measures, totaling 15 parameters for this data. The unrestricted model for the three subscales in the two-level SEM is then

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \Sigma_w = \begin{bmatrix} \sigma_{W11} & & \\ \sigma_{W21} & \sigma_{W22} & \\ \sigma_{W31} & \sigma_{W32} & \sigma_{W33} \end{bmatrix}, \quad \Sigma_B = \begin{bmatrix} \sigma_{B11} & & \\ \sigma_{B21} & \sigma_{B22} & \\ \sigma_{B31} & \sigma_{B32} & \sigma_{B33} \end{bmatrix}. \tag{10.32}$$

Using the truly unrestricted two-level SEM for comparison, the model in Equation 10.30 is, in fact, rejected with $\chi^2(9) = 85.62$, $p < .001$. The M*plus* syntax for this analysis is provided in Appendix C-1.

The next model we consider is nearly identical to Equation 10.30 but relaxes the assumption of equality for the factor loadings:

$$\nu = \begin{bmatrix} 0 \\ \gamma_{100} \\ \gamma_{200} \end{bmatrix}; \quad \Lambda_W = \Lambda_B = \begin{bmatrix} 1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix};$$
$$\alpha = \begin{bmatrix} \gamma_{000} \end{bmatrix}; \ \Theta_W = \sigma^2 \mathbf{I}; \ \Theta_B = \mathbf{0}; \ \Psi_W = \begin{bmatrix} \tau_\pi \end{bmatrix}; \ \Psi_B = \begin{bmatrix} \tau_\beta \end{bmatrix}. \tag{10.33}$$

In this model, we clearly see that the two-level SEM allows for both differences in intercepts (difficulty) and factor loadings (discrimination) across the measured variables. Freeing these two factor loadings results in a significant improvement in model fit relative to the model in Equation 10.31, $\chi^2(2) = 7.61$, $p = .022$. This improvement in fit is not sufficient to result in a good fitting model overall. Relative to the unrestricted model, the model in Equation 10.33 still is rejected: $\chi^2(7) = 78.01$, $p < .001$. The M*plus* syntax for this analysis is provided in Appendix C-2.

We next estimated a model that allowed for the estimation of different factor loadings and residual variances within- and between-groups:

$$\nu = \begin{bmatrix} 0 \\ \gamma_{100} \\ \gamma_{200} \end{bmatrix}; \ \Lambda_W = \begin{bmatrix} 1 \\ \lambda_{W2} \\ \lambda_{W3} \end{bmatrix}; \ \Lambda_B = \begin{bmatrix} 1 \\ \lambda_{B2} \\ \lambda_{B3} \end{bmatrix};$$
$$\alpha = \begin{bmatrix} \gamma_{000} \end{bmatrix}; \ \Theta_W = \sigma_W^2 \mathbf{I}; \ \Theta_B = \sigma_B^2 \mathbf{I}; \ \Psi_W = \begin{bmatrix} \tau_\pi \end{bmatrix}; \ \Psi_B = \begin{bmatrix} \tau_\beta \end{bmatrix}. \tag{10.34}$$

By removing the equality constraints on Λ_W and Λ_B, we are allowing the within-groups relation between the factor and observed variables to differ from the between-groups relation between the factor and the observed variable group means. The addition of residual variance at the group level admits the possibility that not all of the variability in the group means is due to differences in the common factor mean. Some group-level variance is specific to each measured variable. The addition of the three new parameters in this model resulted in a dramatic improvement in model fit relative to the model in Equation 10.34: $\chi^2(3) = 72.07$, $p < .001$. Furthermore, the model in Equation 10.34 could not be rejected by comparison to the unrestricted model: $\chi^2(4) = 5.94$, $p = .204$, indicating that this model adequately recovers the relations between the three observed measures at both the student and school levels. The estimates of the factor loadings and variance components from this model are reported in Table 10.10. The M*plus* syntax is provided in Appendix C-3.

In fact, Table 10.10 shows that although this model fits better than the original HLM, the differences in the estimates are rather minor. The estimated factor loadings are all close to one, and the residual variance for the random intercepts, though significantly different from zero, is small (.061). The within-schools factor variance estimate is trivially larger (3.041, relative to 2.998 in the HLM model), and the between-schools factor variance estimate shows a somewhat larger difference (.718, relative to .683 in the HLM model). Given the high power of the current analyses, these differences were statistically significant but might not be substantively meaningful. In other applications, larger differences between the two approaches could occur.

TABLE 10.10 Parameter Estimates and Standard Errors for Two-Level SEM in Equation (10.34)

Parameter	Estimate	Standard error
λ_{w2}	1.026	.027
λ_{w3}	0.964	.026
$\lambda_{\beta 2}$	0.950	.093
$\lambda_{\beta 3}$	0.932	.092
σ^2_W	2.653	.046
σ^2_B	0.061	.016
τ_α	3.041	.135
τ_β	0.718	.217

Limitations of the Two-Level SEM

Thus, we see that the two-level SEM offers some additional flexibility for fitting hierarchical models with latent variables. Our example illustrated some of these added features but not others, for instance, the ability to model causal relations among latent factors (see Liang & Bentler, 2004, for an example and useful discussion of the full multilevel SEM). The two-level SEM approach does, however, have its own limitations. The commonly used MUML (Muthén maximum likelihood) estimator requires complete data on the observed indicators of the latent factors and makes the assumption that the observed variables are continuous-normal (to construct the pooled within covariance matrix). More recently, some SEM software has introduced a full-information maximum likelihood estimator that can accommodate missing data and/or other scale types. For other scale types, however, numerical integration methods that become infeasible for large models with many random effects are implemented.

A final limitation applies to both the two-level SEM and the HLM approach equally. Generally, both approaches assume that the measured variables differ only in their intercepts (item difficulties) across groups. That is, there are no random slopes (item discriminations) in the models. More ideally, the full range of HLMs presently fit to observed outcome variables also would be available for latent outcome variables. For instance, the prediction of one latent variable, achievement, by another latent variable, peer acceptance, might have a random slope that, in turn, depends on a group-level variable, classroom climate. While some SEM software (e.g., M*plus*; Muthén & Muthén, 2004) now permits the estimation of random slopes in latent variable models, this requires numerical integration, so the number of latent variables or random effects is limited in practice. Bayesian approaches (e.g., Markov Chain Monte Carlo methods; Ansari et al., 2002) may prove more flexible, but these approaches

bring their own difficulties (e.g., long computing times, difficulty determining convergence). Thus, although many advances have been made in the fitting of hierarchical models with latent variables, much work still remains to be done.

CONCLUSION

In this chapter, we presented multilevel measurement models from the HLM, HGLM, and multilevel SEM perspectives. While traditional measurement models, such as CTT and IRT models, do not take into account the dependency of measures within groups, such as schools, we demonstrated the possibilities of modeling such a nested data structure in measurement models, both for continuous and dichotomous measurement indicators. Also, we presented example data analyses to model different classical test theory assumptions to test their fit to the data with continuous measurement indicators, as well as a model with dichotomous measurement indicators that includes a covariate and additional variance-covariance components in the group-level of the model. Although our discussion was limited to a unidimensional case, similar modeling can be employed for multidimensional cases (e.g., Cheong & Raudenbush, 2000; Kamata & Cheong, 2007). Also, our discussion indicated there are many issues that need further improvement, including computational issues for three-level models with item discrimination parameters for categorical measurement indicators and models with random item discrimination parameters. It is our hope that further advancement will be made in these areas.

NOTES

1. A computer generated data set and supplemental document with syntax for HLM and M*plus* can be obtained through the book web site. The data set on the web is similar in design but not the same as the one used in this chapter for data security reasons.
2. These values could be reproduced by formulating a two-level HLM where the level-one units are subscales and the level-two units are students by taking the ratio of the level-one error variance to the total error variance. (For details, see Miyazaki & Skaggs, in press).
3. There are other ways to parameterize the effect of subscales. For example, see Cheong & Raudenbush (2000) and Kamata & Cheong (2007).
4. Some of the assumptions can be relaxed as we will show in the section of Measurement Models by Multivariate 3-level Model.
5. The HLM syntax for this analysis is provided in Appendix A-1.
6. When this formula is applied to the results for a two-level measurement model ignoring the nested data structure, $\alpha = .798$, will be reproduced (see Miyazaki & Skaggs, in press) in HLM output.

7. Note that the average had to be taken because of the different number of student (J_k). If J_ks were constant, the average was unnecessary, as we did in Equation 10.5.
8. In HLM software (Raudenbush, Bryk, Cheong, Congdon, & Du Toit, 2004), the only option available for multivariate hierarchical linear models is full maximum likelihood.
9. As we will see later, this is not truly an unrestricted model from the multilevel confirmatory factor analysis perspective.
10. Thus, this unrestricted model is not completely unrestricted and will be considered again from a Structural Equation Modeling perspective later in this chapter.
11. HLM syntax for this analysis is provided in Appendix A-2. This syntax produces results for the following two models as well.
12. Extensions to models for polytomously scored items also are shown in Rijmen et al. (2003), Shin (2003), and Williams and Beretvas (2006).
13. This equation is expressed as a combined form with simplified subscripts to highlight its equivalency to the Rasch model. HGLM formulation will be presented later in this section
14. See Beretvas and Kamata (2005) and Kamata (2001) for more details about the relationship between HGLM and the Rasch model.
15. The HLM syntax for this analysis is provided in Appendix B-1.
16. See, for example, Holland and Wainer (1993) and Zumbo (1999) for detailed introduction to DIF.
17. The HLM syntax for this analysis is provided in Appendix B-2.
18. For many years, the MUML estimator was the only estimator available in conventional software for estimating two-level SEMs. This estimator is exact if the number of individuals in each cluster is the same (i.e., balanced) but approximate otherwise. Given the more recent addition of true ML estimation to conventional software (even for unbalanced designs), it is hard to imagine an application where MUML would now be preferable to ML.

ACKNOWLEDGEMENTS

Authors started this work when they were involved in the 2004–2005 program on Latent Variable Models in the Social Sciences (LVSS) at the Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC. Authors are thankful for the opportunities and support from the SAMSI.

REFERENCES

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika, 67*, 49–78.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*, 135–167.

Beretvas, S. N., & Kamata, A. (2005). The multilevel measurement model: Introduction to the special issue. *Journal of Applied Measurement, 6*, 247–254.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley & Sons.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing, 6*, 57–79.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods, 5*, 477–495.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*, 529–569.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Goldstein, H. I., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 455–467.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Hox. J. (2005). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

Kamata, A., & Bauer, D. J. (2008). A note on the relationship between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*, 136–193.

Kamata, A. Chaimongkol, S., Genc, E., & Bilir, M. K. (2005, April). *Random-effect differential item functioning across group unites by the hierarchical generalized linear model.* Paper presented at the annual meeting of American Educational Research Association, Montreal, Canada.

Kamata, A., & Cheong, Y. F. (2007). Multilevel Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 271–232). New York: Springer.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage Publications.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Liang, J., Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika, 69*, 101–122.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage Publications.

Lord, F. M., & Novic, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–271). Charlotte, NC: Information Age Publishing.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42*, 215–232.

Miyazaki, Y. (2005). Some links between classical and modern test theory via the two-level hierarchical generarized linear model. *Journal of Applied Measurement, 6*(3), 289–310.

Miyazaki, Y., & Skaggs, G. (in press). Linking classical test theory and two-level hierarchical linear models. *Journal of Applied Measurement.*

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376–398.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus User's Guide* (3rd ed.). Los Angeles, CA: Muthén & Muthén.

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 199–242). Charlotte, NC: Information Age Publishing.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika 69*, 167–190.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & Du Toit, M. (2004). *HLM 6: hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16*(4). 295–330.

Rijmen, F., & Briggs, D. (2004). Multiple person dimensions and latent item predictors. In P. De Boeck, & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach.* (pp. 247–265). New York: Springer.

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

Rijmen, F., Tuerlinckx, F., Meulders, M., Smits, D. J. M., & Balazs, K. (2005). Mixed model estimation methods for the Rasch model. *Journal of Applied Measurement, 6*, 273–288.

Roberts, J. K., & Herrington, R. (2005). Demonstration of software programs for estimating multilevel measurement model parameters. *Journal of Applied Measurement, 6*, 255–272.

Shin, S. (2003). *A polytomous nonlinear mixed model for item analysis.* Unpublished doctoral dissertation, University of Texas at Austin, Austin, TX.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton: Chapman & Hall/CRC.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Thum, Y. M. (1997). Hierarchical linear models for multivariate behavioral data. *Journal of Educational and Behavioral Statistics, 22*, 77–108.

Traub, R. (1994). *Reliability for the Social Sciences.* Thousand Oaks, CA: Sage.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363–381.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*, 22–42.

Zumbo, B. D. (1999). *A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

APPENDIX A

HLM syntax with 3 continuous measurement indicators

1. Univariate Analysis

```
#WHLM CMD FILE FOR hlm3.mdm
nonlin:n
numit:100
stopval:0.0000010000
level1:MATH10_3=INTRCPT1+D2+D3+RANDOM
level2:INTRCPT1=INTRCPT2+random/
level3:INTRCPT2=INTRCPT3+random/
level2:D2=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
level2:D3=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
fixtau2:3
fixtau3:3
accel:5
level1weight:none
level2weight:none
level3weight:none
varianceknown:none
hypoth:n
resfil1:n
resfil2:n
resfil3:n
constrain:N
laplace:N,0
graphgammas:C:\HLM Book Chapter\three level model\grapheq.geq
lvr-beta:n
title:3 level CTT model
output:C:\HLM Book Chapter\three level model\ctt hlm3.txt
fulloutput:n
fishertype:2
```

2. Multivariate Analyses

```
#WHLM CMD FILE FOR mctt_3l.mdm
numit:100
stopval:0.0000010000
level1:MATH10_3=INTRCPT1+D2+D3+RANDOM
level2:INTRCPT1=INTRCPT2+random
level3:INTRCPT2=INTRCPT3+random
level2:D2=INTRCPT2
level3:INTRCPT2=INTRCPT3
level2:D3=INTRCPT2
level3:INTRCPT2=INTRCPT3
fixtau2:3
fixtau3:3
accel:5
hypoth:n
graphgammas:C:\HLM Book Chapter\M10c_3var\MHLM3\grapheq.geq
r_e_model:hetl1var
title:Multivariate 3 level CTT model
output:C:\HLM Book Chapter\M10c_3var\MHLM3\mctt_3l.txt
fulloutput:n
```

APPENDIX B

HLM syntax for 22 dichotomous measurement indicators

1. Unconditional Model

```
#WHLM CMD FILE FOR Ch12_2.mdm
nonlin:binomial
microit:50
macroit:200
stopmicro:0.0000010000
stopmacro:0.0001000000
level1:RESPON=INTRCPT1+I4+I6+I7+I8+I10+I14+I16+I17+I19+I21+I
       25+I26+I28+I29+I30+I31+I33+I35+I36+I38+I39+RANDOM
level2:INTRCPT1=INTRCPT2+random/
level3:INTRCPT2=INTRCPT3+random/
level2:I4=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
level2:I6=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
level2:I7=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
```



```
.
.
.
level2:I39=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
fixsigma2:1.000000
fixtau2:3
fixtau3:3
accel:5
level1weight:none
level2weight:none
level3weight:none
varianceknown:none
hypoth:n
resfil1:n
resfil2:n
resfil3:n
constrain:N
laplace:N,50
graphgammas:F:\Chapter12\Ch12_2a.geq
lvr-beta:n
title:no title
output:F:\Chapter12\Ch12_2a.out
fulloutput:n
fishertype:2
```

2. Model with Level-2 Covariate and Its Random Effect at Level-3

```
#WHLM CMD FILE FOR Ch12_2.mdm
nonlin:binomial
microit:50
macroit:200
stopmicro:0.0000010000
stopmacro:0.0001000000
level1:RESPON=INTRCPT1+I4+I6+I7+I8+I10+I14+I16+I17+I19+I21+I
       25+I26+I28+I29+I30+I31+I33+I35+I36+I38+I39+RANDOM
level2:INTRCPT1=INTRCPT2+LUNCH+random/
level3:INTRCPT2=INTRCPT3+random/
level3:LUNCH=INTRCPT3/
level2:I4=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
level2:I6=INTRCPT2+LUNCH/
level3:INTRCPT2=INTRCPT3/
level3:GROUP=INTRCPT3+random/
```

```
level2:I7=INTRCPT2/
level3:INTRCPT2=INTRCPT3/

.
.
.

level2:I39=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
fixsigma2:1.000000
fixtau2:3
fixtau3:3
accel:5
level1weight:none
level2weight:none
level3weight:none
varianceknown:none
hypoth:n
resfil1:n
resfil2:n
resfil3:n
constrain:N
laplace:N,50
graphgammas:F:\Chapter12\Ch12_2b.geq
lvr-beta:n
title:no title
output:F:\Chapter12\Ch12_2b.out
fulloutput:n
fishertype:2
```

APPENDIX C

M*plus* syntax for 2-level SEM with 3 continuous measurement indicators

1. Model with Unit Factor Loadings

```
TITLE: CTT one factor, two levels, unit loadings,
       homoscedastic;
DATA: FILE IS Ch12_3.dat;
VARIABLE:
NAMES ARE student school lunch i1-i21 area1 area2 area3;
USEVARIABLES school area1 area2 area3;
CLUSTER IS school;
```



```
MISSING IS .;
ANALYSIS:
TYPE = MEANSTRUCTURE TWOLEVEL;
ESTIMATOR = ML;
MODEL:
%BETWEEN%
math_b by area1@1;
math_b by area2@1 (1);
math_b by area3@1 (2);
area1@0 area2@0 area3@0;
[area1*5 area2*5 area3*5];
math_b*.7;
%WITHIN%
math_w by area1@1;
math_w by area2@1 (1);
math_w by area3@1 (2);
math_w*3;
area1*2.7 (3);
area2 (3);
area3 (3);
```

2. Model with Heterogeneous Loadings ($\Lambda_W = \Lambda_B$)

```
TITLE: CTT one factor, two levels, free loadings,
       homoscedastic;
DATA: FILE IS Ch12_3.dat;
VARIABLE:
NAMES ARE student school lunch i1-i21 area1 area2 area3;
USEVARIABLES school area1 area2 area3;
CLUSTER IS school;
MISSING IS .;
ANALYSIS:
TYPE = MEANSTRUCTURE TWOLEVEL;
ESTIMATOR = ML;
MODEL:
%BETWEEN%
math_b by area1@1;
math_b by area2*1 (1);
math_b by area3*1 (2);
area1@0 area2@0 area3@0;
[area1*5 area2*5 area3*5];
math_b*.7;
%WITHIN%
math_w by area1@1;
```

```
math_w by area2*1 (1);
math_w by area3*1 (2);
math_w*3;
area1*2.7 (3);
area2 (3);
area3 (3);
```

3. Model with Heterogeneous Loadings ($\Lambda_W \neq \Lambda_B$)

```
TITLE: CTT one factor, two levels, within and between
       loadings and intercepts;
DATA: FILE IS Ch12_3.dat;
VARIABLE:
NAMES ARE student school lunch i1-i21 area1 area2 area3;
USEVARIABLES school area1 area2 area3;
CLUSTER IS school;
MISSING IS .;
ANALYSIS:
TYPE = MEANSTRUCTURE TWOLEVEL;
ESTIMATOR = ML;
MODEL:
%BETWEEN%
math_b by area1@1;
math_b by area2*1;
math_b by area3*1;
!M10MCB@0 M10MCD@0 M10MCE@0;
area1 (1);
area2 (1);
area3 (1);
[area1*5 area2*5 area3*5];
math_b*.7;
%WITHIN%
math_w by area1@1;
math_w by area2*1;
math_w by area3*1;
math_w*3;
area1*2.7 (2);
area2 (2);
area3 (2);
```

PART IV

MASTERING THE TECHNIQUE