The Global Issues and Trends in Educational Statistics: Merging Psychometric and Statistical Models

Akihito Kamata

ABSTRACT

One recent trend in educational statistics is treatments of psychometric models as special cases of statistical models. For example, some classical test theory and item response theory models have been shown that they can be modeled under a framework of multilevel modeling, as well as under the framework of factor analysis and structural equation modeling. A major benefit of these integrated approaches is that it allows educational researchers to easily extend traditional psychometric models to more complex ones. In this paper, several aspects of such generalization of psychometric models are reviewed.

Multilevel Item Response Model

Multilevel modeling has become a popular data analysis framework in educational research (e.g., Raudenbush & Bryk, 2002). Typical application of multilevel modeling is in studying nested data, such as students nested within school, and studying repeated measures, such as modeling growth. Interestingly, a 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. 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 an extension of a regression model to multiple levels, the flexibility of MMM offers the opportunity to incorporate covariates and their interaction effects to a psychometric model.

HGLM Approach for Multilevel Item Response Theory Model

A desirable statistical approach to modeling categorical item responses is a generalized linear model (GLM) extension of the multilevel linear model, such as the hierarchical generalized linear model (HGLM - Raudenbush & Bryk, 2002); specifically one using the logit link. In this section, we assume that the responses are scored dichotomously. However, extension to models for polytomously scored items are straight forward, and shown in such as Rijmen et al. (2003), Shin (2003), and Williams and Beretvas (2006).

Let Y_{ijk} = 1 if the *i*th response is correct for student *j* of school k and Y_{ijk} = 0 otherwise, and μ_{ijk} be the probability of Y_{ijk} = 1. This probability varies randomly across students. However, conditioning on this probability, we have $Y_{ijk} \mid \mu_{ijk} \sim$ *Bernoulli* with $E(Y_{ijk} \mid \mu_{ijk}) = \mu_{ijk}$, and $var(Y_{ijk} \mid \mu_{ijk}) = \mu_{ijk}(1 - \mu_{ijk})$. Then, a multilevel measurement model can be written for the example data set as:

$$\text{logit}(\mu_{ijk}) = \gamma_i + r_{jk} + u_k , \qquad (1)$$

where γ_i is the effect of item i. Student ability variation within school, $r_{jk} \sim N(0, \sigma^2_{r_k})$ and the school mean ability variation $u_k \sim N(0, \sigma^2_u)$ 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 difference from the Rasch model is that this 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. In fact, equation (1) can be simplified to a 2-level model by not considering the level-3 variation,

$$\text{logit}(\mu_{ij}) = \pi_i + r_j . \qquad (2)$$

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. See Kamata (2001) and Beretvas and Kamata (2005) for more details about the relationship between HGLM and the Rasch model.

Models in equations (1) and (2) are hierarchical generalized linear models based on a logit link function; 3-level HGLM for (1) and a 2-level HGLM for (2). 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 (2) 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 type of model can be modeled in the framework of HLM by 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, rather than specifying a reference item (e.g., Cheong & Raudenbush, 2000).

Parameter estimation can be accomplished by any of several different methods, including the penalized quasi likelihood (PQL), Laplace approximation, Gaussian numerical integration of log-likelihood, and fully Baysian Markov Chain Monte Carlo (MCMC) methods. Rijmen et al. (2005) provides a comparison of these estimation methods and

demonstrated that all estimators performed equally well in reasonable conditions. For the interested reader, Roberts and Herrington (2005) demonstrate how to set up and analyze data for these models in several different software packages.

Equation (1) is fit by a 3-level HGLM for an example data set. The example data set includes the simulated item response data for 22 dichotomously scored items with a total of 3,312 students (N) from 30 schools (K). In the illustrative data, there are 22 items. Therefore, the level-1 equation is the item level model and can be written

$$\text{logit}(\mu_{ijk}) = \pi_{0jk} + \sum_{q=1}^{21} \pi_{qjk} D_{qijk} , \quad (3)$$

where D_{qijk} is the qth indicator variable that takes a value of 1 if $q = i$ for item i. There is no error term in (3) because it is absorbed by the link function. One item is used as a reference item, and item difficulties are assessed relative to the reference item. Thus, $q = 1, \ldots, 21$, rather than up to 22. Then, the level-2 equations are

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + r_{0jk} \\ \pi_{qjk} &= \beta_{q0k} , \end{aligned} \quad (4)$$

where $q = 1, \ldots, 21$. Slopes are not random because item difficulties are assumed to be equal across individual students. The level-3 equations are

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + r_{00k} \\ \beta_{q0k} &= \gamma_{q00} . \end{aligned} \quad (5)$$

Slopes are not randome because 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. The results of this model are presented in Table 1.

The difficulty of item 22 (the reference item) was estimated to be .344 (γ_{000}) in logit, indicating it is .344 higher than the mean ability. Other values indicated in Table 1a are differences in their difficulties compared to item 22. For example, item 1 is more difficult than item 22 by .178, thus its difficulty is .344 + .178 = .522. On the other hand, item 2 is easier than item 22 by .408, so its difficulty is .344 - .403 = - .059. Notice that se(γ_{i00}) for item 22 is the standard error for the difficulty, while the standard errors for the remaining estimated parameters are standard errors for the difference in difficulty from that of item 22. For this model, the variances of the student abilities are provided in the bottom panel of Table 1b. The within-school variance was estimated as sigma-squared = var(r_{0jk}) = .694, while the between-school variance was estimated as tau = var (u_{00k}) = .162. This implies that the intra class correlation in latent mathematics ability is $u_{00k} / (r_{0jk} + u_{00k})$ = .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 1 Results of example data analysis by HGLM - unconditional model.

a. Fixed effects

Item	γ_{i00}	se(γ_{i00})	Item	γ_{i00}	se(γ_{i00})
1	.178	.054	12	-.006	.053
2	-.403	.053	13	-.508	.053
3	.023	.053	14	.399	.054
4	.028	.053	15	.377	.054
5	.061	.053	16	-.485	.053
6	-1.342	.055	17	.507	.055
7	.883	.057	18	-.977	.054
8	-.555	.053	19	-.201	.053
9	-.265	.053	20	-.874	.054
10	-.785	.053	21	1.061	.058
11	-1.266	.055	22*	.344	.084

* 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}.

b. Random effects

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

Model Extensions with a Covariate

The multilevel IRT models can be easily extended to include covariates and additional variance and covariance components. For example, let's assume we have an additional person-level predictor (X_{jk}) in the 3-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 (3) as the level-1 model. The level-2 equations become

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + \beta_{01k}X_{jk} + r_{0jk} \\ \pi_{qjk} &= \beta_{q0k} + \beta_{q1k}X_{jk}, \end{aligned} \tag{6}$$

where q = 1, ..., 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 are also interested in the random variation of the person by item interaction effect across the level-3 units, the level-3 equations become

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + u_{00k} \\ \beta_{01k} &= \gamma_{010} \\ \beta_{q0k} &= \gamma_{q00} \\ \beta_{q1k} &= \gamma_{q10} + u_{q1k}, \end{aligned} \tag{7}$$

where q = 1, ..., 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}), variance and covariance components var(r_{0jk}), var(u_{00k}), var(u_{q1k}) 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. In fact, this 3-level formulation of the model is equivalent to the "random-effect DIF model" presented by Cheong (2006) and Kamata et al. (2005). In this context, one's interest is to estimate 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, a dummy variable that may represent student characteristics, such as enrollment in a free or subsidized lunch program, is used. Fifty-seven percent of the students in the sample were coded to 1, as opposed to 0. The model was fit by the HGLM option in the HLM software, 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 7) for identification reasons. Our preliminary data exploration indicated the magnitude of the DIF for the third item in the measurement subscale was near zero (γ_{i10} = .0001 for this item). Therefore, its effect was constrained to be zero, and the parameter was dropped from the model. Accordingly, var(μ_{i1k}) was also constrained to be zero for this item.

Seven items displayed significant DIF (DIF estimates were larger than twice their standard errors), and they are indicated by asterisks in Table 2. All of them had negative values, indicating students who had been coded 1 on the covariate had significantly lower odds of correct answers for the indicated items, given the same level of ability. For item 20 in the data analysis subscale, for example, the odds of correct answer for students with code 1 was only 70% (exp[-.351] = .704) of the odds of success for students who had code 0. For the seven items that displayed significant DIF, random effects were further estimated in a separate model. Six of these items displayed significant variability, indicating statistically meaningful variations of DIF across schools. For example, the estimate of var(μ_{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 cor (u_{00k}, μ_{i1k}) in Table 2, we see that the correlations are all positive among the seven items experiencing significant DIF. Positive correlations indicate that DIF were higher for schools with higher mean performance, because these seven items had negative DIF magnitudes, indicating students with code 1 had lower odds of correct answer, given the same level of ability. However, since 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 that was close to zero for schools with higher mean performance.

Table 2 Estimates of fixed and random effects of the interaction effect.

Item	Fixed effects		Random effects	
	γ_{i10}	se	var(μ_{i1k})	cor(u_{00k}, μ_{i1k})
1	-.247*	.113	.072†	.134
2	.000†	-		
3	.024	.133		
4	-.297	.110	.059†	.436
5	-.149	.111		
6	-.066	.107		
7	-.307*	.117	.078†	.573
8	.139	.128		
9	-.155	.103		
10	-.284*	.117	.144†	.453
11	-.069	.108		
12	.007	.110		
13	-.072	.105		
14	-.417*	.105	.010†	.476
15	-.199	.113		
16	.182	.107		
17	-.264*	.114	.091†	.643
18	-.040	.122		
19	-.037	.142		
20	-.351*	.128	.230†	.499
21	.111	.124		
22	.000†	-		

* Magnitudes are greater than twice the standard errors.

† Magnitudes are significant at α = .05 based on chi-square test.

One advantage of formulating these models in the HGLM framework as compared to the traditional (single-level) 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. 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 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 is typically done in confirmatory factor analysis and two-parameter item response models. The need for this modeling flexibility is suggested by many empirical applications of linear factor models and item response models where the factor loadings are seen to differ across the observed measures.

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 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 (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 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 et al. (2003) and Rijmen and Briggs (2004) provide an example of such an approach using a non-linear mixed model. According to their approach, a 2-parameter logistic IRT model is modeled by treating a logit of the probability of the response as a linear function. We can assume that the distribution of the latent trait to be 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 2-level models. (This does not preclude the inclusion of level-3 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-3 model, such as $\mathrm{var}(u_{00k})$, $\mathrm{var}(u_{q1k})$, $\mathrm{cov}(u_{00k}, u_{q1k})$ in Equation (7).

An additional limitation of the 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., Raudenbush, Rowan & Kang, 1991; Cheong & Raudenbush, 2000; Kamata & Cheong, 2007). However, the structure applied to the covariance matrix among these latent variables is often 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.

Given these limitations of the HGLM approach to the design and analysis of measurement models, one alternative is the Generalized, Linear, Latent and Mixed Model

(GLLAMM) (Skrondal & Rabe-Hesketh, 2004). This modeling 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 (6) and (7), along with discrimination parameters. In this example, however, we had 8 random effects at level 3 of the model (7 random DIF and 1 school level variance of latent abilities), and this number of random effects unfortunately makes numerical integration computationally impractical.

Another alternative approach that has a similar flexibility is the 2-level structural equation model (SEM). It is this approach that I discuss in greater detail in the next two sections.

Item Response Theory Model as a Factor Analytic Model

The relationship between binary factor analysis (FA) and the two-parameter IRT model has been clarified by Takane and de Leeuw (1987) and McDonald (1999). Among those, Kamata and Bauer (2008) pointed out different parameterizations of a unidimensional binary FA model are available and provided general conversion formulas to convert FA model parameters to IRT parameters under several different parameterizations of the binary FA model.

Factor Analytic Model for Binary Variables

The key assumption to the binary FA is that there is a continuous underlying latent response, denoted y_i^*, that is an additive combination of the common factor and item-specific residual. A one-factor model for the latent response variable can thus be written as

$$y_i^* = \nu_i + \lambda_i \xi + \varepsilon_i, \tag{8a}$$

where ν_i is the intercept, λ_i is the factor loading, the latent factor score for a particular person is ξ, and ε_i is the residual for item i (for compactness, no person subscript is included). In addition, the residuals ε_i are typically assumed to be normally distributed, but a logistic distribution can also be considered.

A threshold part of the model is then added to accommodate the dichotomous nature of the observed response, y_i:

$$y_i = \begin{cases} 1 \text{ if } y_i^* \geq \tau_i \\ 0 \text{ if } y_i^* < \tau_i. \end{cases} \quad (8b)$$

where, τ_i is the threshold for item i. Since the intercepts ν_i and thresholds τ_i are not jointly identified, the intercepts are typically assumed to be zero, consequently we will not consider the intercepts further. It is important to note that the scale of both predictor (ξ) and outcome (y_i^*) are not fixed, since they are unobserved latent variables. Therefore, the scales of parameters (λ_i and τ_i) does not inherently exist, either. Parameterizations for this model can be classified by the scaling of (a) the continuous latent response variables, y_i^* and (b) the continuos common factor ξ.

As one approach, the variance of is arbitrarily (but without a loss of generality) constrained to be 1.0 for all items. If we do so, the residual variance of y_i^*, var(ε_i) or var ($y_i^*|\xi$), is obtained as var(ε_i) = 1 - λ_i^2 var(ξ). This parameterization has been common in binary FA (Muthén & Asparouhov, 2002; Millsap & Yun-Tein, 2004). Since the marginal distribution of the continuous latent trait score (y_i^*) is standardized, Kamata and Bauer (2008) referred this parameterization to as the Marginal parameterization.

Another strategy is to constrain the residuals ε_i to have a fixed variance, such as 1. With this arrangement, the variances of y_i^* are obtained as var (y_i^*) = λ_i^2 var(ξ)+1. This parameterization is less commonly used with binary FA, but it is closer to the conventional two-parameter IRT parameterization. Since the conditional distribution of the continuous latent trait score ($y_i^*|\xi$) is standardized, Kamata and Bauer (2008) referred this parameterization to as the Conditional parameterization.

Two common scaling conventions are to choose a reference indicator or to standardize the common factor. In the former approach, the threshold and factor loading of one item are fixed to zero and one, respectively (e.g., $\tau_1 = 0$ and $\lambda_1 = 1$). These constraints allow the mean and the variance of ξ to be freely estimated. In the latter approach, the mean and the variance of ξ are constrained to be zero and one [e.g., $E(\xi) = 0$ and $\text{var}(\xi) = 1$]. Consequently, all λ_i and τ_i are freely estimated.

By the combination of the two types of scaling choices mentioned above, four different parameterizations are obtained. They are summarized in Table 1 of Kamata and Bauer (2008).

Relation of Binary FA to Two-Parameter IRT

The two-parameter IRT model can be written

$$p_i(y_i = 1 \mid \xi) = f(\alpha_i \xi + \beta_i), \tag{9}$$

where α_i is the slope parameter and β_i is the intercept parameter for item i, ξ is the latent trait score for a specific person, and f is a cumulative distribution function (CDF), chosen to be either a normal or logistic CDF. Takane and de Leeuw (1987) derived the relationship between parameters from the *standardized* FA and IRT models, that is, when assuming zero mean and unit variance for the latent factor. For unidimensional models of the kind considered here, their formulas reduce to

$$\alpha_i = \frac{\lambda_i}{q_i} \text{ and } \beta_i = \frac{-\tau_i}{q_i}, \tag{10}$$

where q_i is $\sqrt{\text{var}(\varepsilon_i)}$. Although not explicitly stated by Takane & de Leeuw, these formulas make clear that the Conditional-Standardized binary factor model parameterization is in fact equivalent to the IRT model, except for the reversal of sign for the threshold parameter τ_i relative to β_i (given that $q_i = 1$ in this parameterization). Under the Marginal-Standardized parameterization, $q_i = \sqrt{1-\lambda_i^2}$, paralleling formulas given in a number of

other references (e.g., McDonald, 1999; p.259). These results are very useful for understanding the relationship between FA and IRT models. Takane and de Leeuw's derivation is not, however, directly applicable to factor models using the reference indicator convention for scaling the latent factor.

Kamata and Bauer (2008) recently derived and demonstrated a set of more general formula. They are:

$$\alpha_i = \frac{\lambda_i \sqrt{\text{var}(\xi)}}{\sqrt{\text{var}(\varepsilon_i)}} \text{ and } \beta_i = \frac{-[\tau_i - \lambda_i E(\xi)]}{\sqrt{\text{var}(\varepsilon_i)}}. \quad (11)$$

Then, Kamata and Bauer rewrote these general formulas into 4 sets of conversion formulas for the four different parameterizations by substituting the appropriate values for these formulas. Their results of substitutions are also presented here in Table 3.

Table 3 Conversion formulas for 4 factor-analysis parameterizations

	Reference Indicator	**Standardized Indicator**
Marginal	$\alpha_i = \frac{\lambda_i \sqrt{V(\xi)}}{\sqrt{1-\lambda_i^2 V(\xi)}}$ $\beta_i = \frac{-[\tau_i - \lambda_i E(\xi)]}{\sqrt{1-\lambda_i^2 V(\xi)}}$	$\alpha_i = \frac{\lambda_i}{\sqrt{1-\lambda_i^2}}$ $\beta_i = \frac{-\tau_i}{\sqrt{1-\lambda_i^2}}$
Conditional	$\alpha_i = \lambda_i \sqrt{V(\xi)}$ $\beta_i = -[\tau_i - \lambda_i E(\xi)]$	$\alpha_i = \lambda_i$ $\beta_i = -\tau_i$

Why is it important?

As Kamata and Bauer (2008) indicated, the choice between parameterizations is arbitrary, because the results obtained under one specification can be translated directly into the results under the other parameterizations. However, we should be mindful that which parameterization we choose may influence interpretations when extending to more complex models.

One advantage of the Marginal parameterization is that the total variance of y_i^* is held constant as new predictors (either latent factors or item covariates) are added to the equation for that item. As the explained variance in y_i^* goes up, the residual variance, var (ε_i), goes down, just as in a typical linear model. In contrast, if new predictors are added to the equation in the Conditional parameterization, the residual variance, var(ε_i), continues to be held constant and as the explained variance goes up, the total variance of y_i^* must also go up. As a consequence, all of the coefficients of earlier predictors (e.g, factor loadings) must be adjusted to the new scale (regardless of whether the old and new predictors, e.g., factors, are correlated). These adjustments due to the changing scale of y_i^* are non-intuitive for researchers more familiar with linear models and can make interpretation of the coefficients more difficult. On the other hand, the Conditional parameterization has other advantages. As discussed by Millsap and Yun-Tein (2004) and Muthen and Asparouhov (2002), the Conditional parameterization may be preferable when extending the binary FA to make comparisons across multiple groups or over multiple time points. The reason is that in the Conditional parameterization, var(ε_i) is a parameter in the model that is directly accessible to the researcher, whereas in the Marginal parameterization it is an remainder involving several other model parameters. Having direct access to var(ε_i) can be important if there is reason to believe this variance may differ over groups or over time, in which case it can be standardized in one group/time and estimated in another (provided that certain other constraints are implemented to identify the model).

Regarding the choice between a reference indicator and the standardized factor, standardizing the factor is a simple and elegant approach that many factor analysts use regularly. Further, when used in concert with the Marginal parameterization and a normal distribution for ε_i, the thresholds fall on a standard normal curve, easing the translation of these estimates into marginal proportions for the observed responses. Other factor analysts, however, may prefer to use a reference indicator so that the latent factor will be in the same metric as the chosen indicator. While this is a compelling rationale when the indicators are continuous, for binary items the metric of the latent response variables is ultimately arbitrary (for instance, depending on the choice of Conditional or Marginal

parameterizations), so the reference indicator approach does not seem to provide a similar interpretational advantage in this case. One advantage of the reference indicator approach is that the mean and variance of the latent factor can be estimated, which may facilitate across-group or over-time comparisons. Another possible advantage is that if there is a particular item that one wishes to compare other items to, this item can be chosen as the referent. The threshold parameters estimated for the remaining items will then reflect differences in difficulty or severity relative to the referent item.

2-Level Structural Equation Model

The binary FA model can be extended as a more general framework for a multilevel IRT modeling, namely, a 2-level structural equation model with categorical indicators. The 2-level SEM is different from the traditional single-level SEM in assuming that multiple individuals are sampled from each of many groups in the population. The two-level factor model with categorical indicators can be written as

$$\mathbf{y}^{*}_{pg} = \mathbf{\Lambda}_{W}\boldsymbol{\theta}_{pg} + \boldsymbol{\varepsilon}_{pg}, \tag{12}$$

which is a linear regression of the vector of I unobserved latent response variables $\mathbf{y}^{*}_{pg}$, on the latent variables $\boldsymbol{\theta}_{pg}$ for person p in group g. $\mathbf{y}^{*}_{pg}$ is an I x 1 vector that contains latent response scores to I items, while $\boldsymbol{\theta}_{pg}$ is a K x 1 vector that contains latent scores for K latent factors. As such, $\mathbf{\Lambda}_{W}$ are factor loadings (I x K matrix) and $\boldsymbol{\varepsilon}_{pg}$ are residuals (I x 1 vector), where the W subscript indicates "within-groups". In a unidimensional IRT application, for example, K = 1. Consequently, both $\mathbf{\Lambda}_{W}$ and $\boldsymbol{\varepsilon}_{pg}$ are I x 1 vectors, where I is the number of items in the test. Observed dichotomous response y_{ipg} is defined such that

$$y_{ipg} = 1, \text{ if } y^{*}_{ipg} \geq \tau_{i}, \text{ and} \tag{13}$$

$$y_{ipg} = 0, \text{ if } y^{*}_{ipg} < \tau_{i}.$$

Here, τ_{i} is the threshold for item i. Within groups, the latent factors are assumed to be distributed with mean vector α and covariance matrix $\mathbf{\Psi}_{W}$. Similarly, for polytomously scored items with scoring categories ranging from 0 to M,

$$
\begin{aligned}
y_{ipg} &= M \text{ , if } y^{*}_{ipg} \geq \tau_{iM} \text{ ,} \\
y_{ipg} &= M-1 \text{, if } \tau_{i(M-1)} \leq y^{*}_{ipg} \leq \tau_{iM} \text{ ,} \\
&\vdots \\
y_{ipg} &= 1 \text{ , if } \tau_{i1} \leq y^{*}_{ipg} \leq \tau_{i2} \text{ , and} \\
y_{ipg} &= 0 \text{ , if } y^{*}_{ipg} < \tau_{i1} \text{ .}
\end{aligned}
\tag{14}
$$

The residuals $\boldsymbol{\varepsilon}_{pg}$ are assumed to be distributed with means of zero and covariance matrix $\boldsymbol{\Sigma}_{W}$. Usually $\boldsymbol{\Sigma}_{W}$ is assumed to be diagonal, reflecting independent residuals or local independence. If errors are distributed as the logistic distribution, the model is known as the logistic model, and this will provide the basis of equivalency to logistic item response models. If we have $\boldsymbol{\theta}_{pg}$ as a 1 x 1 scalar (i.e., only one latent trait) and M = 2 (i.e., dichotomously scored items), the model is equivalent to the 2PL IRT model. On the other hand, if residuals are normally distributed, the model is known as the normal ogive model. One important assumption with this approach is that these covariance matrices are homogeneous across all groups, which will result in identical covariance structures between groups. For any given group *j*, the within-group covariance matrix is given by essentially the same equation as the single-level SEM,

$$
V(\mathbf{y}^{*})_{W} = \boldsymbol{\Lambda}_{W} \boldsymbol{\Psi}_{W} \boldsymbol{\Lambda}'_{W} + \boldsymbol{\Sigma}_{W} \text{ .} \tag{15}
$$

On the other hand, the structural model of the SEM can be written as

$$
\boldsymbol{\theta}_{pg} = \boldsymbol{\alpha}_{g} + \mathbf{B}_{g} \boldsymbol{\theta}_{pg} + \boldsymbol{\Gamma}_{g} \mathbf{x}_{pg} + \boldsymbol{\zeta}_{pg} \text{ ,} \tag{16}
$$

where latent variables are regressed on other latent variables, as well as on some observed covariates $\mathbf{x}$. The intercepts are given by $\boldsymbol{\alpha}_{g}$, slopes for latent predictors are $\mathbf{B}_{g}$, and slopes for observed covariates are $\boldsymbol{\Gamma}_{g}$. The residuals are assumed to be normally distributed with means of zero and K x K covariance matrix $\boldsymbol{\psi}$. If there is no latent variable regression in an SEM (for instance, in a confirmatory factor analysis), the intercepts $\boldsymbol{\alpha}_{j}$, are simply interpreted as factor means (which are typically constrained to be 0) and $\boldsymbol{\psi}$ is the covariance matrix of the latent factors.

The key difference between the 2-level SEM and the standard single-level SEM involves the additional component of variability due to nested data structure. The multilevel SEM imposes an additional factor structure on the covariance matrix (Ansari, Jedidi & Dube, 2002; Goldstein & McDonald, 1988; McDonald & Goldstein, 1989; Muthén, 1995; Muthén & Satorra, 1994). The resulting covariance structure at between-group level is

$$V(y^*)_B = \Lambda_B \psi_B \Lambda_B' + \Sigma_B, \text{ and}$$
$$V(y^* \mid \mathbf{x})_B = \Lambda_B (\mathbf{I} - \mathbf{B}_B)^{-1} \psi_B (\mathbf{I} - \mathbf{B}_B)'^{-1} \Lambda_B' + \Sigma_B. \quad (17)$$

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 (Equations 15 - 17), the differential subscripting of the matrices W or B indicates that the parameter estimates or even the factor structure of the model can differ between the two parts of the model. Assuming that only one latent trait is measured and regression of the latent trait is only on some observed covariates, the above general SEM will be reduced to be the multilevel IRT model, described earlier.

Traditionally, parameter estimation for this type of model has relied on the weighted least squares (WLS) methods with a tetrachoric or polychoric correlation matrix, which differs from IRT estimation tradition, where a full information maximum likelihood has been a tradition. Also, the scaling of parameters will be different from the parameter scale of IRT if the weighted least square is employed, which requires appropriate transformation of parameters (Kamata & Bauer, 2008) as mentioned earlier in this paper. More recently, a true full information maximum likelihood estimator has become available in several general SEM software programs, which is consistent with the IRT tradition. Also, the MCMC has been shown to be effective especially when the number of random effects becomes large (e.g., Fox & Glas, 2003; Fox, 2005, 2007; Chaimongkol, 2005; Vaughn, 2006).

By this approach, the same data analysis presented by the HGLM approach can be performed. Of course, a benefit of the multilevel SEM approach is that the model can

incorporate variant item discriminations, while HGLM approach must restrict item discriminations the same for all items.

Conclusion

In this paper a multilevel item response model was presented both from the HGLM and multilevel SEM perspectives. While traditional measurement models, such as IRT models, do not take into account of dependency of measures within groups, such as schools, it was demonstrated the possibilities of modeling such a nested data structure in measurement models. Although the discussion was limited to a unidimensional case, similar modeling can be employed for multidimensional cases (e.g., Cheong & Raudenbush, 2000; Kamata & Cheong, 2007). Also, there are many issues that need further improvement, including computational issues for 3-level models with item discrimination parameters for categorical measurement indicators and models with random item discrimination parameters. It is hoped that further advancement will be made in these areas.

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.

Chaimongkol, S. (2005). *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective.* Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

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 childrens problem behaviors. *Psychological Methods, 5,* 477-495.

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

Fox, J. -P., & Glass, C. A. W. (2003). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika,* 66, 269-286.

Fox, J. -P. (2005). Multilevel IRT Model Assessment. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.) *New developments in categorical dat analysis for the social and behavioral sciences* (pp.227-252). Mahwah, NJ: Lawrence Erlbaum Associates.

Fox, J. -P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software,* 20(5), 1-16.

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

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, 136-153.

Kamata, A. Chaimongkol, S., Genc, E., & Bilir, M. K. (2005). *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, April 2005.

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

McDonald, R. P. (1999). *Test theory: Unified treatment.* Lawrence Erlbaum Associates.

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.

Millsap, R. E. & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research,* 39, 479-515.

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

Muthén, B. O. & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus.* Mplus Web Note No. 4, at http://www.statmodel.com/mplus/examples/webnote.html.

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), *Sociological Methodology 1995,* 216-316.

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., and 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.

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach.* Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

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