Research Report

On the Estimation of Hierarchical Latent Linear Models for Large Scale Assessments

Deping Li

Andreas Oranje

Research & Development

December 2006
RR-06-37

On the Estimation of Hierarchical Latent Linear Models for Large Scale Assessments

Deping Li and Andreas Oranje
ETS, Princeton, NJ

December 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

A hierarchical latent regression model is suggested to estimate nested and nonnested relationships in complex samples such as found in the National Assessment of Educational Progress (NAEP). The proposed model aims at improving both parameters and variance estimates via a two-level hierarchical linear model. This model falls naturally within the set of models used in most large scale surveys, in that all of them are special cases of the hierarchical latent regression model. The model parameter estimates are obtained via the expectation-maximization (EM) algorithm. An example with NAEP data is presented and results of parameter estimation and standard errors are compared with results from operational procedures of NAEP.

Key words: Hierarchical linear models, latent regression models, maximum likelihood estimates, EM algorithm, item response theory, NAEP

Acknowledgments

The authors would like to thank Dr. Shelby Haberman and Dr. John Donoghue for their suggestions, reviews, and discussions with the authors for a variety of issues involved in this project. We thank Catherine McClellan, Tamás Antal, and Dan Eignor for their reviews and comments.

1 Introduction

Hierarchical linear modeling (HLM) is a well-known framework particularly appropriate for complex stratified and clustered samples with balanced or unbalanced designs and varying degrees of missing data patterns (Raudenbush & Bryk, 2002). This family of models has proven to be useful especially for the analysis of educational data from complicated designs involving survey sampling. Examples of cognitive educational surveys with complex designs are the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Studies (TIMSS). In these programs students are sampled from schools, and schools are sampled from participating states or countries following a stratification scheme. Stratification in this context means that the sample is drawn from each of several relatively homogenous strata with respect to characteristics such as median household income and level of urbanization. Furthermore, blocks of items are systematically sampled across students to assess a large amount of information in a relatively short amount of student testing time. Hence, students do not answer all test questions, and relatively little information is obtained from individual students. Justification for this design stems from the reporting level of these programs, being proficiency estimates of student populations (e.g., male students, students in a particular state or country).

The prevalent operational model used to estimate student population proficiencies is a latent regression model (Allen, Donoghue, & Schoeps, 2001; Mislevy, 1984, 1985). Population groups indicators are regressed onto a latent proficiency, which is characterized by a collection of item response theory (IRT) models (Birnbaum, 1968; Lord, 1980; van der Linden & Hambleton, 1997). Suppose that the ability of student i in subscale t is denoted as θ_{it}, where $i = 1, \cdots, n$ and $t = 1, \cdots, p$. For example, in the NAEP mathematics assessment p equals 5, corresponding to the subscales: (a) Numerical Properties and Operations; (b) Measurement; (c) Geometry; (d) Data Analysis and Probability; and (e) Algebra. Furthermore, θ is a latent variable and, hence, the latent regression assumes the following form:

$$\theta_{it} = \boldsymbol{\gamma}_t \boldsymbol{x}_i + \varepsilon_i, \tag{1}$$

where $\boldsymbol{\gamma}_t$ is a vector of Q regression coefficients for scale t, that is, $\boldsymbol{\gamma}_t = (\gamma_{1t}, \cdots, \gamma_{Qt})'$. Furthermore, $\boldsymbol{x}_i$ is a vector of Q population groups indicators and ε_i is a residual term, assumed to be normal distributed. Student latent proficiencies $\boldsymbol{\theta}_t = (\theta_{t1}, \cdots, \theta_{tN})'$ can be inferred from student item responses $\boldsymbol{Y}_t = (\boldsymbol{y}_{1t}, \cdots, \boldsymbol{y}_{Nt})$ through item response theory. The complete latent

regression model assuming a normal distribution for θ_i given $\mu_{it} = \boldsymbol{\gamma}_t{}' \boldsymbol{x}_i$ can be estimated through the following likelihood:

$$L_t = \prod_{i=1}^{N} \left(\left(\prod_{h=1}^{M} P(y_{iht} | \theta_{it}, \boldsymbol{\beta}_h) \right) \phi(\theta_{it} | \mu_{it}, \sigma_t^2) \right)^{w_i} \tag{2}$$

where P denotes the probability of observing response y_{iht} given latent ability θ_i and IRT item parameters $\boldsymbol{\beta}$. Furthermore, ϕ denotes a normal density with variance σ_t^2. The product operators imply that items are assumed to be independently observed and that students are independently observed. This is obviously not true in a complex sample, and therefore post-hoc operations are conducted in NAEP, using jackknife repeated replications (JRR), to derive appropriate standard errors. Several alternatives have been proposed.

Wilson and Adams (1995) and Adams, Wilson, and Wu (1997) have formulated a multilevel model to concurrently estimate latent regression and IRT parameters. Also, Raudenbush, Fotiu, and Cheong (1999) have applied a multilevel model to the plausible values produced in the NAEP program, which are imputations from the posterior ability distribution. Furthermore, Johnson (2002) and Johnson and Jenkins (2005) have developed a Bayesian framework for multilevel IRT with NAEP data that include a Markov chain Monte Carlo estimation procedure and also concurrent estimation of item and population parameters. Finally, Aitkin and Aitkin (2005) have used generalized mixed models to estimate four-level models with NAEP-type data.

These alternatives have in common that they work quite well with very small samples and small regression models. However, larger data sets corresponding to typical NAEP samples become rapidly intractable to compute. The only exception is Raudenbush et al. (1999), although the extent to which the multilevel structure is captured in the plausible values is questionable. In this study, a less ambitious approach has been taken to estimate a random effects parameter across schools in the population model. When students are selected from the same school, observations of school and teacher characteristics are quite likely to be related, violating the assumption of independent observations that is required in NAEP's latent regression models. More importantly, as Raudenbush and Bryk (2002) pointed out, students in the same school share values on many more variables, some of which are not observed, which means that the variables tend to disappear into the residual term of the linear model, causing correlated disturbances.

There are several other clustering levels in educational surveys like NAEP, such as states or countries, classrooms, and primary sampling units (PSUs). These are ignored in this study for

several reasons. For classrooms, sampling often occurs across subjects (e.g., reading, mathematics) and across students. In practice, not enough data is available to support such detailed estimation. For states and PSUs, the situation is more complex. In state-to-state samples, PSUs coincide with schools and a separate model is applied to each state after a common measurement scale is determined. In other words, the state level of clustering is implicit. For national samples, geographic location is more likely to be an important clustering variable, although stratification may be applied at this level as well, hence reducing the effect of clustering.

The first section of this paper will introduce the structure of a latent hierarchical linear model that is appropriate for NAEP data, followed by a discussion of model parameter estimation via the expectation-maximization (EM) algorithm. Also, estimation of standard errors is discussed and the model is extended to the multivariate case. The univariate case is illustrated by an application of a two-level model to NAEP 2004 age 17 long-term trend mathematics data, and the resulting parameter estimates are compared with parameters from the current operational NAEP approach. The paper concludes with a discussion of these results.

2 Hierarchical Latent Regression Model

2.1 Hierarchical Linear Models

The primary focus of this study is hierarchical modeling with latent variables. To facilitate the discussion of this model, we will use a simple two-level latent regression model with predictors only on the first level. This model can be considered a special case of a two-level hierarchical latent regression model (HLRM), which may or may not be more appropriate for the problem at hand. Extension of this model to a general HLRM is also straightforward. Since the discussion of hierarchical models involves sampling clusters, the notation of clusters will be added to indicate hierarchical relations and a nested data structure. Hence, there are n_j students nested within school j for $j = 1, \cdots, J$, and their proficiencies θ_{1jt} to $\theta_{n_j jt}$ are likely to be positively correlated, since they share a common curriculum and instructional experience. Furthermore, let $\boldsymbol{x}_{ij}$ be the vector of Q population groups indicators with an additional cluster index, that is, $\boldsymbol{x}_{ij} = (X_{ij1}, X_{ij2}, \cdots, X_{ijQ})$ for student i in school j. Correspondingly, the regression effects for school j and subscale t is $\boldsymbol{\gamma}_{jt} = (\gamma_{jt1}, \cdots, \gamma_{jtQ})'$. The item responses for student i in school j on subscale t is denoted by $\boldsymbol{y}_{ijt}$. A two-level model describes linear relations between the latent proficiency and the population groups indicators where regression coefficients are random effects

across school j. The Level 1 model is:

$$\theta_{ijt} = \boldsymbol{x}_{ij}\boldsymbol{\gamma}_{jt} + \varepsilon_{ijt}. \tag{3}$$

The residual term ε_{ijt} is assumed to be $\varepsilon_{ijt} \sim N(0, \sigma_t^2)$. The Level 2 model is:

$$\boldsymbol{\gamma}_{jt} = \boldsymbol{\gamma}_t + \boldsymbol{u}_{jt}. \tag{4}$$

The coefficients $\boldsymbol{\gamma}_t$ in the above model (4) without the second-level notation implies that the overall regression effects (a vector of Q regression effect parameters (i.e.,$\boldsymbol{\gamma}_t = (\gamma_{t1}, \cdots, \gamma_{tQ})'$)) are fixed, that is, they do not vary across sampling units.

With the second-level model, predictors can be added to model the random regression effects. Furthermore, this model can be made quite flexible by estimating some regression effects with second-level predictors and some without and possibly by allowing a subset of regression parameters $\boldsymbol{\gamma}_{jt}$ to be considered random effects $\boldsymbol{u}_{jt}$ and others to be considered fixed effects. In this paper, a simple unconditional model on the second level (i.e., the model in (4) without predictors) is used primarily for demonstration purposes. Moreover, in the simple model, $\hat{\gamma}_t$ is an overall estimated mean of regression effects across all clusters, taking into account school or cluster random effects. Hence, it is expected that the HLRM model will help in the estimation of relations between student abilities and the group indicators. The random effects parameter $\boldsymbol{u}_{jt}$ is assumed to be $\boldsymbol{u}_{jt} \sim N(\boldsymbol{0}, \boldsymbol{T}_t)$ for $j = 1, 2, \cdots J$ and $t = 1, \cdots, p$. A rationale for treating cluster regression effects as random effects is that NAEP does not report individual school effects but rather group proficiency for the population of students. Because participating schools as well as their students are sampled from this population, there is primarily interest in the variance component due to the complex sample and not in specific school or student effects.

Substituting the model in (4) in (3) gives the combined model

$$\theta_{ijt} = \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t + \boldsymbol{x}_{ij}\boldsymbol{u}_{jt} + \varepsilon_{ijt}. \tag{5}$$

The marginal variance of θ_{ijt} is $\boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x}'_{ij} + \sigma^2$, where $\boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x}'_{ij}$ is the variance component due to random effects $\boldsymbol{u}_{jt}$, attributable to the variation across selected schools. Also, σ_t^2 depicts the variation among students within schools. In this model, variation is decomposed into a school- and a student-level component.

2.1.1 *Multivariate Case*

In the multivariate case, student proficiency $\boldsymbol{\theta}_{ij}$ is a p-dimensional vector (i.e., $\boldsymbol{\theta}_{ij} \in \Re^p$ and $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \cdots, \theta_{ijp})'$). Regression effects on p subscales can be represented as a matrix $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1 | \boldsymbol{\gamma}_2 | \cdots | \boldsymbol{\gamma}_p)$, where each column is a vector of regression effects for a particular subscale or dimension. The random school effects are represented as $\boldsymbol{U}_j = (\boldsymbol{u}_{j1} | \boldsymbol{u}_{j2} | \cdots | \boldsymbol{u}_{jp})$, and the residuals are a column vector of p subscale residuals $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \cdots, \varepsilon_{ijp})'$. Subsequently, the combined multivariate model is a hierarchical linear regression among p-dimensions and can be written as

$$\boldsymbol{\theta}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\Gamma} + \boldsymbol{x}_{ij}\boldsymbol{U}_j + \boldsymbol{\varepsilon}_{ij}. \tag{6}$$

Alternatively, this model can be written by using matrix notation that reflects only the second-level and/or subscale indices. For example, in the univariate case, for subscale $t = 1, \cdots, p$, the proficiencies of all students within a second-level unit (e.g., in school j) can be expressed as $\boldsymbol{\theta}_{jt} = (\theta_{1jt}, \cdots, \theta_{n_j jt})'$, and for group indicators $\boldsymbol{X}_j = (\boldsymbol{x}'_{1j}, \cdots, \boldsymbol{x}'_{n_j j})'$. Furthermore, the residual terms are $\boldsymbol{\varepsilon}_{jt} = (\varepsilon_{1jt}, \cdots, \varepsilon_{n_j jt})'$, leading to a model for school j

$$\boldsymbol{\theta}_{jt} = \boldsymbol{X}_j\boldsymbol{\gamma}_t + \boldsymbol{X}_j\boldsymbol{u}_{jt} + \boldsymbol{\varepsilon}_{jt}. \tag{7}$$

This notation is used frequently throughout this report. In addition, the model is also expressed as

$$\theta_{ijt} = \sum_{q=1}^{Q} X_{ijq}\gamma_{qt} + \sum_{q=1}^{Q} X_{ijq}u_{qjt} + \varepsilon_{ijt} \tag{8}$$

when discussing computational details.

2.2 *Estimation of HLRM Parameters*

In NAEP, the sample is stratified both at the school- and at the student-level, resulting in unequal sampling probabilities. Let w_{ij} be the sampling weight for student i in school j. While there is much discussion about the use of sampling weights within the HLM framework (e.g., Asparouhov, 2005), they are included here in HLRM models in order to capture differences between sampling and population distributions regardless of the complex nature of the sample. In the following sections, the marginal likelihood estimates (MML) of the model parameters will be discussed as part of an EM algorithm. For subscale $t = 1, \cdots, p$, the *log* likelihood function L over

all individual students and test item responses is

$$
\begin{aligned}
L &= log\left(\prod_{j=1}^{J}\prod_{i=1}^{n_j} P(\boldsymbol{y}_{ijt}|\boldsymbol{x}_{ij}\boldsymbol{\gamma}_t, \boldsymbol{x}_{ij}\boldsymbol{T}_t\boldsymbol{x'}_{ij} + \sigma_t^2)^{w_{ij}}\right) \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij} log\left(\int P(\boldsymbol{y}_{ijt}|\theta)\phi(\theta|\boldsymbol{x}_{ij}\boldsymbol{\gamma}_t, \boldsymbol{x}_{ij}\boldsymbol{T}_t\boldsymbol{x}'_{ij} + \sigma_t^2)d\theta\right). \qquad (9)
\end{aligned}
$$

In the maximization step, the parameters $\boldsymbol{\gamma}_t$, σ_t^2, and $\boldsymbol{T}_t$ given the data $(\boldsymbol{u}_{jt}, \boldsymbol{y}_{ijt})$ for $i = 1, \cdots, n_j$, $j = 1, \cdots, J$ and $t = 1, \cdots, p$ can be estimated by (see also the appendix)

$$
\hat{\boldsymbol{\gamma}}_t = \left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\boldsymbol{x'}_{ij}\boldsymbol{x}_{ij}\right)^{-1} \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\boldsymbol{x'}_{ij}(\tilde{\theta}_{ijt} - \boldsymbol{u}_{jt}\boldsymbol{x}_{ij}), \qquad (10)
$$

$$
\hat{\sigma}_t^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\tilde{\sigma}_{ijt}^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}(\tilde{\theta}_{ijt} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}}, \qquad (11)
$$

$$
\hat{\boldsymbol{T}}_t = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{u}_{jt}\boldsymbol{u'}_{jt}, \qquad (12)
$$

where $\tilde{\theta}_{ijt}$ in (10) is the posterior mean of student proficiencies and $\tilde{\sigma}_{ijt}^2$ in (11) is the posterior variance. In the expectation step (E-step), these moments can be found following

$$
\tilde{\theta}_{ijt} = \int \theta_{ijt} p(\theta_{ijt}|\boldsymbol{y}_{ijt}) d\theta_{ijt}, \qquad (13)
$$

and

$$
\tilde{\sigma}_{ijt}^2 = \int (\theta_{ijt} - \tilde{\theta}_{ijt})^2 p(\theta_{ijt}|\boldsymbol{y}_{ijt}) d\theta_{ijt}, \qquad (14)
$$

where the posterior density $p(\theta_{ijt}|\boldsymbol{Y})$ can be expressed following Bayes theorem as

$$
p(\theta_{ijt}|\boldsymbol{Y}) = \frac{p(\boldsymbol{y}_{ijt}|\theta_{ijt})\phi(\theta_{ijt}|\boldsymbol{x}_{ij}\boldsymbol{\gamma}_t, \boldsymbol{x}_{ij}\boldsymbol{T}_t\boldsymbol{x'}_{ij} + \sigma_t^2)}{\int p(\boldsymbol{y}_{ijt}|\theta_{ijt})\phi(\theta_{ijt}|\boldsymbol{x}_{ij}\boldsymbol{\gamma}_t, \boldsymbol{x}_{ij}\boldsymbol{T}_t\boldsymbol{x'}_{ij} + \sigma_t^2)d\theta_{ijt}}. \qquad (15)
$$

Furthermore, $\boldsymbol{T}_t$ is a covariance matrix of random school effects

$$
\boldsymbol{T}_t = \begin{pmatrix} \tau_{t11} & \tau_{t12} & \cdots & \tau_{t1Q} \\ \tau_{t21} & \tau_{t22} & \cdots & \tau_{t2Q} \\ \cdots & \cdots & \cdots & \cdots \\ \tau_{tQ1} & \tau_{tQ2} & \cdots & \tau_{tQQ} \end{pmatrix},
$$

where the diagonal elements τ_{tqq} indicate the variance of random regression effects of γ_{qtj} for

$q = 1, \cdots, Q$ and $t = 1, \cdots, p$ across all schools $j = 1, \cdots, J$. $\tau_{tqq'}$ and the off-diagonal elements indicate the covariance between two random effects γ_{qtj} and $\gamma_{q'tj}$ for $q, q' = 1, \cdots, Q$.

Extension of the maximization step to the multivariate case is straightforward. The estimates of $\boldsymbol{\Gamma}, \boldsymbol{\Sigma}$, and $\boldsymbol{T} = Var(\boldsymbol{U}_j)$ for $t = 1, \cdots, p$ and $j = 1, \cdots, J$ can be estimated by

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{t}} = \left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \boldsymbol{x}'_{ij} \boldsymbol{x}_{ij} \right)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \boldsymbol{x}'_{ij} (\tilde{\theta}_{ijt} - \boldsymbol{u}_{jt} \boldsymbol{x}_{ij}), \tag{16}$$

for each subscale $t = 1, \cdots, p$. Collecting the estimates of $\hat{\gamma}_t$ will yield $\hat{\boldsymbol{\Gamma}} = [\hat{\gamma}_1, \cdots, \hat{\gamma}_p]$:

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \tilde{\boldsymbol{\Sigma}}_{ij} + \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \tilde{\boldsymbol{\varepsilon}}'_{ij} \tilde{\boldsymbol{\varepsilon}}_{ij}}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij}}, \tag{17}$$

$$\hat{\boldsymbol{T}} = \frac{1}{J} \sum_{j=1}^{J} \boldsymbol{U}'_j \boldsymbol{U}_j, \tag{18}$$

where $\tilde{\boldsymbol{\varepsilon}}_{ij} = (\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij} \boldsymbol{\Gamma} - \boldsymbol{x}_j \boldsymbol{U}_j)$ in (17). Furthermore, the posterior mean is now a vector of posterior means, and $\tilde{\boldsymbol{\Sigma}}_{ij}$ is the posterior variance-covariance matrix of $\boldsymbol{\theta}_{ij}$. The covariance matrix $\hat{\boldsymbol{T}}$ in (18) has dimension $pQ \times pQ$. The diagonal block matrix in $\hat{\boldsymbol{T}}$ is $\hat{\boldsymbol{T}}_{tt}$ for $t = 1, \cdots, p$ and the off-diagonal block matrix $\hat{\boldsymbol{T}}_{ts}$ is the covariance matrix of school random effects between subscale t and subscale s, for $s, t = 1, \cdots, p$. Hence, the matrix $\hat{\boldsymbol{T}}$ can be further written in terms of the variance and covariance of random school effects among p subscales as the following block matrix:

$$\hat{\boldsymbol{T}} = \begin{pmatrix} \hat{\boldsymbol{T}}_{11} & \hat{\boldsymbol{T}}_{12} & \cdots & \hat{\boldsymbol{T}}_{1p} \\ \hat{\boldsymbol{T}}_{21} & \hat{\boldsymbol{T}}_{22} & \cdots & \hat{\boldsymbol{T}}_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\boldsymbol{T}}_{p1} & \hat{\boldsymbol{T}}_{p2} & \cdots & \hat{\boldsymbol{T}}_{pp} \end{pmatrix},$$

with

$$\hat{\boldsymbol{T}}_{ts} = \frac{1}{j} \sum_{j=1}^{J} \boldsymbol{u}_{jt} \boldsymbol{u}'_{js}, \tag{19}$$

and

$$\hat{\boldsymbol{T}}_{ts} = \hat{\boldsymbol{T}}'_{st} \tag{20}$$

for $t \neq s$ and $t, s = 1, \cdots, p$, denoting the covariance matrices of random school effects between subscale t and s. In the multivariate case, the covariance of the effects between any two subscales t and s for $t \neq s$, $t, s = 1, \cdots, p$, is a $Q \times Q$ matrix with elements

$$\hat{\boldsymbol{T}}_{ts} = \begin{pmatrix} \hat{\tau}_{ts,11} & \hat{\tau}_{ts,12} & \cdots & \hat{\tau}_{ts,1Q} \\ \hat{\tau}_{ts,21} & \hat{\tau}_{ts,22} & \cdots & \hat{\tau}_{ts,2Q} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\tau}_{ts,Q1} & \hat{\tau}_{ts,Q2} & \cdots & \hat{\tau}_{ts,QQ} \end{pmatrix}.$$

where $\hat{\boldsymbol{T}}_{tt}$ is defined in (18), representing the covariance matrix of school random effects for subscale t. Therefore, there are p variance matrices $\hat{\boldsymbol{T}}_t$ and $\frac{p(p-1)}{2}$ covariance matrices $\hat{\boldsymbol{T}}_{ts}$ resulting in $\frac{p^2+p}{2}$ matrices to be estimated in order to obtain $\hat{\boldsymbol{T}}$ in the multivariate case.

2.2.1 Conditional Moments for the Univariate Case

Individual student proficiency $\boldsymbol{\Theta}$ (i.e., $\boldsymbol{\theta}_{jt}$, θ_{ijt}) and the random school effect $\boldsymbol{U}$ (or $\boldsymbol{u}_{jt}$) are unobserved quantities. However, conditional expectations can be obtained given observed response data $\boldsymbol{Y}$ (i.e., $\boldsymbol{y}_{jt}$, $\boldsymbol{y}_{ijt}$) and parameter estimates from a previous EM cycle $\hat{\boldsymbol{\gamma}}_t, \hat{\sigma}_t^2, \hat{\boldsymbol{T}}_t$. Specifically, student latent proficiencies $\boldsymbol{\theta}_{jt}$ and school effects $\boldsymbol{u}_{jt}$ are treated as missing data for $j = 1, \cdots, J$ and $t = 1, \cdots, p$. If $(\boldsymbol{\theta}_{jt}, \boldsymbol{u}_{jt})$ was observed, finding the MML estimates would be straightforward. For notational convenience, let $\boldsymbol{\Omega} = (\boldsymbol{X}_j, \boldsymbol{y}_j, \boldsymbol{\sigma}_t^2, \boldsymbol{\gamma}_t, \boldsymbol{T}_t)$, including the observed responses $\boldsymbol{y}_j$ and group indicators $\boldsymbol{X}_j$, as well as previous parameter estimates from a previous cycle $\boldsymbol{\gamma}_t, \boldsymbol{\sigma}_t^2, \boldsymbol{T}_t$.

The posterior mean and variance for student abilities can be obtained through numerical integration using (13) and (14). Hence, to complete the E-step of this algorithm, an expression for the conditional expectations for $E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ and $E\left[(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}\right]$ to estimate σ_t^2, and $E(\boldsymbol{u}_{jt}\boldsymbol{u}'_{jt}|\boldsymbol{\Omega})$ to estimate $\boldsymbol{T}_t$ are needed. The question is how these conditional expectations can be found.

For the term $(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2$ in (11) it follows that

$$(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2 = (\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t)^2 - 2(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t)\boldsymbol{x}_{ij}\boldsymbol{u}_{jt} + (\boldsymbol{x}_{ij}\boldsymbol{u}_{ij})^2. \tag{21}$$

Therefore, finding the conditional expectation for $E[(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}]$ only requires the conditional expectations $E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega}))$, $E[(\boldsymbol{x}_{ij}\boldsymbol{u}_{ij})^2|\boldsymbol{\Omega}]$, and $E(\boldsymbol{u}_{jt}\boldsymbol{u}'_{jt}|\boldsymbol{\Omega})$. Because

$$E(\boldsymbol{u}_{jt}\boldsymbol{u}'_{jt}|\boldsymbol{\Omega}) \;\; = \;\; E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})' + \mathit{Var}(\boldsymbol{u}_{jt}|\boldsymbol{\Omega}), \tag{22}$$

the evaluation of $E(\boldsymbol{u}_{jt}\boldsymbol{u}'_{jt}|\boldsymbol{\Omega})$ relies on the conditional expectation $E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega}))$ and variance $\mathit{Var}(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ of $\boldsymbol{u}_{jt}$. Therefore, in the E-step the conditional expectations and variances of missing data $\boldsymbol{u}_{jt}$, as well as the conditional expectation of the quadratic term $(\boldsymbol{x}_{ij}\boldsymbol{u}_{ij})^2$ given $\boldsymbol{\Omega}$, need to be evaluated before computing parameter estimates for each M cycle. That is, each E-step

requires the evaluation of $E(\boldsymbol{\theta}_{jt}|\boldsymbol{\Omega})$, $Var(\boldsymbol{\theta}_{jt}|\boldsymbol{\Omega})$, $E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$, $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$, and $E[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}]$. They are the elements necessary for the computation of expected sufficient statistics for $\boldsymbol{\gamma}_t, \sigma_t^2, \boldsymbol{T}_t$, provided that the complete data were observed. Substituting the expected complete data sufficient statistics into formulas in (10), (11), and (12) in the M-step yields improved estimates in terms of the likelihood. Repeating the E- and M-steps until convergence is achieved yields maximum likelihood estimates (Dempster, Laird, & Rubin, 1977).

Let $\tilde{\boldsymbol{\theta}}_{jt} = (\tilde{\theta}_{ijt}, \cdots, \tilde{\theta}_{in_jt})$ and $Var(\boldsymbol{\theta}_{jt}|\boldsymbol{\Omega}) = \tilde{\boldsymbol{\Sigma}}_j$, denoting a diagonal matrix with elements $Var(\theta_{1jt}|\boldsymbol{Y}), \cdots, Var(\theta_{n_jjt}|\boldsymbol{Y})$. The conditional mean $E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$, the conditional variance $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ and the conditional expectation for the quadratic form $E[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}]$ depend on the joint distribution of $\boldsymbol{u}_{tj}$ and latent abilities $\boldsymbol{\theta}_{jt}$. The joint distribution of $(\boldsymbol{\theta}_j, \boldsymbol{u}_{jt})$ given $\boldsymbol{X}_j, \boldsymbol{\gamma}_t, \boldsymbol{T}_t, \sigma_t^2$ is assumed to be multivariate normal with mean vector $(\boldsymbol{X}_j\boldsymbol{\gamma}_t, \boldsymbol{0})$ and covariance matrix

$$Cov(\boldsymbol{\theta}_{jt}, \boldsymbol{u}_{jt}) = \begin{pmatrix} \sigma_t^2\boldsymbol{I} + \boldsymbol{X}_j\boldsymbol{T}_t\boldsymbol{X}_j' & \boldsymbol{X}_j\boldsymbol{T}_t \\ \boldsymbol{T}'_t\boldsymbol{X}'_j & \boldsymbol{T}_t \end{pmatrix}.$$

Following the proof of Raudenbush and Bryk (2002, pp. 442–443), the conditional expectation is given by

$$\begin{aligned} E(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_{jt}) &= \boldsymbol{T}_t\boldsymbol{X}'_j(\sigma^2\boldsymbol{I} + \boldsymbol{X}_j\boldsymbol{T}_t\boldsymbol{X}'_j)^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{X}_j\boldsymbol{\gamma}_t) \\ &= \boldsymbol{C}_{jt}^{-1}\boldsymbol{X}'_j(\boldsymbol{\theta}_{jt} - \boldsymbol{X}_j\boldsymbol{\gamma}_t), \end{aligned} \tag{23}$$

where $\boldsymbol{I}$ in (23) indicates an identity matrix of size j for $j = 1, \cdots, J$, and

$$\boldsymbol{C}_{jt} = \boldsymbol{X}'_j\boldsymbol{X}_j + \sigma^2\boldsymbol{T}_t^{-1}. \tag{24}$$

The conditional variance of $\boldsymbol{u}_{jt}$ given $\boldsymbol{\theta}_j$ can be expressed as

$$\begin{aligned} Var(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_j) &= \boldsymbol{T}_t - \boldsymbol{T}_t\boldsymbol{X}'_j(\sigma_t^2\boldsymbol{I} + \boldsymbol{X}_j\boldsymbol{T}_t\boldsymbol{X}'_j)^{-1}\boldsymbol{X}_j\boldsymbol{T}_t \\ &= \boldsymbol{C}_{jt}^{-1}\sigma_t^2. \end{aligned} \tag{25}$$

Denote the conditional expectation $E(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_{jt})$ as $\bar{\boldsymbol{u}}_{jt}$, which can be evaluated through (23) for each school. The values of $\bar{\boldsymbol{u}}_{jt}$ and $Var(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_{jt})$ can be obtained prior to observing item response

data $\boldsymbol{Y}$. The conditional expectation of $\tilde{\boldsymbol{u}}_{jt}$ given $\boldsymbol{\Omega}$ can be expressed as

$$\begin{aligned}
\tilde{\boldsymbol{u}}_{jt} &= \int \boldsymbol{u} \left(\int P(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{Y}) P(\boldsymbol{\theta}|\boldsymbol{Y}, \sigma_t^2, \boldsymbol{\gamma}_t, \boldsymbol{T}_t) d\boldsymbol{\theta} \right) d\boldsymbol{u} \\
&= \int \int \boldsymbol{u} P(\boldsymbol{u}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{Y}, \sigma_t^2, \boldsymbol{\gamma}_t, \boldsymbol{T}_t) d\boldsymbol{u} d\boldsymbol{\theta} \\
&= \int \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j (\boldsymbol{\theta}_{jt} - \boldsymbol{X}_j \boldsymbol{\gamma}_t) P(\boldsymbol{\theta}|\boldsymbol{Y}, \sigma_t^2, \boldsymbol{\gamma}_t, \boldsymbol{T}_t) d\boldsymbol{\theta} \\
&= \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j (\tilde{\boldsymbol{\theta}}_{jt} - \boldsymbol{X}_j \boldsymbol{\gamma}_t). \qquad (26)
\end{aligned}$$

Also, the posterior mean $\tilde{\boldsymbol{u}}_{\boldsymbol{jt}}$ in (26), $E(\sum_{j=1}^{J} \boldsymbol{u}_{jt} \boldsymbol{u}'_{jt})$ can be expressed as

$$E\left(\sum_{j=1}^{J} \boldsymbol{u}_{jt} \boldsymbol{u}'_{jt} \right) = \sum_{j=1}^{J} \tilde{\boldsymbol{u}}_{jt} \tilde{\boldsymbol{u}}'_{jt} + \sum_{j=1}^{J} Var(\boldsymbol{u}_{jt}|\boldsymbol{Y}, \sigma_t^2, \boldsymbol{\gamma}_t, \boldsymbol{T}_t), \qquad (27)$$

where $Var(\boldsymbol{u}_{jt}|\boldsymbol{X}_j, \boldsymbol{Y}, \sigma_t^2, \boldsymbol{\gamma}_t, \boldsymbol{T}_t)$ is the conditional variance for each random school effect. Write $Var(\boldsymbol{u}_{jt}|\boldsymbol{X}_j, \boldsymbol{Y}, \sigma_t^2, \boldsymbol{\gamma}_t, \boldsymbol{T}_t)$ as $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ for simplicity, which can be integrated and expressed in terms of posterior variance of $\boldsymbol{\theta}_{jt}$, denoted as $\tilde{\boldsymbol{\Sigma}}_j$. This is a diagonal matrix with elements equal to the posterior variance for each student within a school or cluster:

$$\begin{aligned}
Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega}) &= \int (\boldsymbol{u}_{jt} - \tilde{\boldsymbol{u}}_{jt})(\boldsymbol{u}_{jt} - \tilde{\boldsymbol{u}}_{jt})' \left(\int P(\boldsymbol{u}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\Omega}) d\boldsymbol{\theta} \right) d\boldsymbol{u} \\
&= \int \int [(\boldsymbol{u}_{jt} - \bar{\boldsymbol{u}}_{jt})(\boldsymbol{u}_{jt} - \bar{\boldsymbol{u}}_{jt})' + (\bar{\boldsymbol{u}}_{jt} - \tilde{\boldsymbol{u}}_{jt})(\bar{\boldsymbol{u}}_{jt} - \tilde{\boldsymbol{u}}_{jt})'] P(\boldsymbol{u}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\Omega}) d\boldsymbol{u} d\boldsymbol{\theta} \\
&= Var(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_j) + \int \int [(\bar{\boldsymbol{u}}_{jt} - \tilde{\boldsymbol{u}}_{jt})(\bar{\boldsymbol{u}}_{jt} - \tilde{\boldsymbol{u}}_{jt})'] P(\boldsymbol{u}|\theta) P(\boldsymbol{\theta}|\boldsymbol{\Omega}) d\boldsymbol{u} d\boldsymbol{\theta} \\
&= \boldsymbol{C}_{jt}^{-1} \sigma_t^2 + \left[\boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j \right] \tilde{\boldsymbol{\Sigma}}_j \left[\boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j \right]'. \qquad (28)
\end{aligned}$$

To evaluate the conditional expectation $E[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}]$, first write it as a quadratic form, $E[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}] = E(\boldsymbol{u}'_{jt}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$. By a theorem of the quadratic expectation (e.g., Stapleton, 1995, p.51),

$$\begin{aligned}
E(\boldsymbol{u}'_{jt}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\boldsymbol{u}_{jt}|\boldsymbol{\Omega}) &= trace[\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\, Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})] + \tilde{\boldsymbol{u}}'_{jt}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{jt} \\
&= \boldsymbol{x}_{ij}\, Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})\boldsymbol{x}'_{ij} + \tilde{\boldsymbol{u}}'_{jt}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{jt}. \qquad (29)
\end{aligned}$$

Combining the expressions of $\tilde{\boldsymbol{u}}_{jt}$ in (26), $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ in (28), and the quadratic term $E[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}]$ in (29) and substituting these into (21) yields $E[(\tilde{\boldsymbol{\theta}}_{ijt} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}]$, which can be written as

$$\begin{aligned}
E[(\tilde{\boldsymbol{\theta}}_{ijt} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}] = (\tilde{\boldsymbol{\theta}}_{ijt} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t)^2 - 2(\tilde{\boldsymbol{\theta}}_{ijt} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t)\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{jt} + \\
\boldsymbol{x}_{ij}\, Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})\boldsymbol{x}'_{ij} + \tilde{\boldsymbol{u}}'_{jt}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{jt}. \qquad (30)
\end{aligned}$$

The expressions of conditional expectation and variance for $\boldsymbol{u}_{jt}$ can be used to complete the computation of parameter estimates for each EM cycle. Note that the estimation of the regression parameters $\boldsymbol{\gamma}_t$, the covariance matrix of the population proficiency distribution σ_t^2, and the covariance matrix $\boldsymbol{T}_t$ of random effects in the M-step only depend on the evaluation of the conditional mean and variance of $\tilde{\theta}_{ijt}$ in the E-step, because the conditional expectations $E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ and $E[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega})]$ and the conditional variance $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ can be written as a function of the conditional moments of θ_{jt}. Therefore, no additional numerical integration is needed to compute $E(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ and $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$, and (15), (13), and (14) are used to obtain these quantities. In this case, a straightforward Simpson rule numerical quadrature integration is used to carry out the computations.

2.2.2 Summary for Univariate Case

The EM algorithm for maximum likelihood parameter estimation in HLRM models for the univariate case can be summarized as follows. The $(r+1)^{th}$ M-step can be completed by computing

$$\boldsymbol{\gamma}_{r+1} = \left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\right)^{-1} \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\boldsymbol{x}'_{ij}(\tilde{\theta}_{ijt} - \boldsymbol{u}_{jt}\boldsymbol{x}_{ij}), \tag{31}$$

$$\sigma^2_{r+1} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\tilde{\sigma}^2_{ijt} + \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}E[(\tilde{\theta}_{ijt} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}_r]}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}}, \tag{32}$$

$$\hat{\boldsymbol{T}}_t = \frac{1}{J}\sum_{j=1}^{J}\tilde{\boldsymbol{u}}_{jt}\tilde{\boldsymbol{u}}'_{jt} + \frac{1}{J}\sum_{j=1}^{J} Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega}_r). \tag{33}$$

The E-step can be completed by computing the posterior moments from (13) and (14) and, subsequently computing $\tilde{\boldsymbol{u}}_{jt}$ using (26), $E[(\tilde{\theta}_{ijt} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})^2|\boldsymbol{\Omega}_r]$ using (30) and $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega}_r)$ using (28). Furthermore, $\boldsymbol{\Omega}_r$ is $\boldsymbol{\Omega}$ at iteration r. The E- and M-steps are alternated until convergence is achieved.

2.2.3 Conditional Moments for the Multivariate Case

It is straightforward to extend the results above to the multivariate case, in which the multivariate conditional expectation $E[(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma} - \boldsymbol{x}_{ij}\boldsymbol{U}_j)'(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma} - \boldsymbol{x}_{ij}\boldsymbol{U}_j)|\boldsymbol{\Omega}]$

needs to be evaluated. Accordingly, $\boldsymbol{\Omega}$ is in this case $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \boldsymbol{T})$. Denote the product $(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma} - \boldsymbol{x}_{ij}\boldsymbol{U}_j)'(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma} - \boldsymbol{x}_{ij}\boldsymbol{U}_j)$ as $\boldsymbol{A}$. Then

$$\boldsymbol{A} = (\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma})'(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma}) - (\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma})'\boldsymbol{x}_{ij}\boldsymbol{U}_j - \boldsymbol{U}_j'\boldsymbol{x}'_{ij}(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\Gamma}) + \boldsymbol{U}_j'\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\boldsymbol{U}_j. \quad (34)$$

The conditional mean $E[\boldsymbol{U}_j|\boldsymbol{\Omega}] = \tilde{\boldsymbol{U}}_j$ is simply the collection of school means for each subscale (i.e., $\tilde{\boldsymbol{U}}_j = [\tilde{\boldsymbol{u}}_{j1}|\tilde{\boldsymbol{u}}_{j2}|\cdots|\tilde{\boldsymbol{u}}_{jp}]$). The posterior variance $Var(\boldsymbol{U}_j|\boldsymbol{\Omega})$ can also be viewed as a collection of posterior variances and covariances of random school effects among subscales. That is, the posterior covariance matrix $Var(\boldsymbol{U}_j|\boldsymbol{\Omega})$ takes on the following form:

$$Var(\boldsymbol{U}_j|\boldsymbol{\Omega}) = \begin{pmatrix} \tilde{\boldsymbol{T}}_{11} & \tilde{\boldsymbol{T}}_{12} & \cdots & \tilde{\boldsymbol{T}}_{1p} \\ \tilde{\boldsymbol{T}}_{21} & \tilde{\boldsymbol{T}}_{22} & \cdots & \tilde{\boldsymbol{T}}_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \tilde{\boldsymbol{T}}_{p1} & \tilde{\boldsymbol{T}}_{p2} & \cdots & \tilde{\boldsymbol{T}}_{pp} \end{pmatrix}.$$

The multivariate conditional expectation and subscale specific variance matrices of random effects can be obtained through (18). Hence, in the multivariate case $\frac{p(p-1)}{2}$ additional conditional covariance matrices $\tilde{\boldsymbol{T}}_{ts} = cov\left[(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|\boldsymbol{\Omega}\right]$ between two subscales for $t, s = 1, \cdots, p$ have to be computed, as well as the $E\left[(\boldsymbol{U}_j'\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\boldsymbol{U}_j)|\boldsymbol{\Omega}\right]$. Furthermore, $\boldsymbol{A}$ contains the $p \times p$ matrix $\boldsymbol{U}_j'\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\boldsymbol{U}_j$. The diagonal elements can be found by (29) following a quadratic form, while the off-diagonal elements can be found by

$$\begin{aligned} E[\boldsymbol{u}'_{js}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\boldsymbol{u}_{jt}|\boldsymbol{\Omega}] &= cov[\boldsymbol{u}'_{js}\boldsymbol{x}'_{ij}, \boldsymbol{x}_{ij}\boldsymbol{u}_{jt}|\boldsymbol{\Omega}] + \tilde{\boldsymbol{u}}'_{js}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{jt} \\ &= \boldsymbol{x}_{ij}\,Cov[\boldsymbol{u}_{jt}, \boldsymbol{u}_{js}|\boldsymbol{\Omega}]\boldsymbol{x}'_{ij} + \tilde{\boldsymbol{u}}'_{js}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{jt} \\ &= \boldsymbol{x}_{ij}\tilde{\boldsymbol{T}}_{st}\boldsymbol{x}'_{ij} + \tilde{\boldsymbol{u}}'_{js}\boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{jt}, \end{aligned} \quad (35)$$

where $\tilde{\boldsymbol{T}}_{st}$ is the posterior covariance between subscale s and subscale t, for $s, t = 1, \cdots p$.

In (29) and (35), two conditional covariances appear: $cov[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt}, \boldsymbol{x}_{ij}\boldsymbol{u}_{jt})|\boldsymbol{\Omega}]$ and $cov[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt}, \boldsymbol{x}_{ij}\boldsymbol{u}_{js})|\boldsymbol{\Omega}]$. The conditional variance $\tilde{\tau}_t$ of $\boldsymbol{u}_{jt}$ for subscale t given $\boldsymbol{\Omega}$ can be evaluated by (28), while the expression for $cov[(\boldsymbol{x}_{ij}\boldsymbol{u}_{jt}, \boldsymbol{x}_{ij}\boldsymbol{u}_{js})|\boldsymbol{\Omega}]$ will be more complicated to obtain. Primarily, $\tilde{T}_{st}$ has to be computed.

First, examine the joint distribution of $\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}, \boldsymbol{u}_{jt}$, and $\boldsymbol{u}_{js}$. The joint distribution of student proficiencies within school j and between two subscales $\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}$ is assumed normal with mean

vector $(\boldsymbol{X}_j\boldsymbol{\gamma}_t, \boldsymbol{X}_j\boldsymbol{\gamma}_s)$, and variance matrix,

$$\boldsymbol{M} = Cov(\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}) = \begin{pmatrix} {\sigma^2}_t\boldsymbol{I} + \boldsymbol{X}_j\boldsymbol{T}_t\boldsymbol{X}'_j & {\sigma^2}_{st}\boldsymbol{I} + \boldsymbol{X}_j\boldsymbol{T}_{ts}\boldsymbol{X}'_j \\ {\sigma^2}_{st}\boldsymbol{I} + \boldsymbol{X}_j{\boldsymbol{T}'}_{ts}\boldsymbol{X}'_j & {\sigma^2}_s\boldsymbol{I} + \boldsymbol{X}_j\boldsymbol{T}_s\boldsymbol{X}'_j \end{pmatrix},$$

where $\boldsymbol{T}_{st} = Cov(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})$, the covariance matrix for random effects between the two subscales. Similarly, the joint distribution of school random effects between two subscales $\boldsymbol{u}_{jt} and \boldsymbol{u}_{js}$ is assumed normal with mean vector $(\mathbf{0}, \mathbf{0})$ and covariance matrix

$$\boldsymbol{G} = Cov(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js}) = \begin{pmatrix} \boldsymbol{T}_t & \boldsymbol{T}_{ts} \\ {\boldsymbol{T}'}_{ts} & \boldsymbol{T}_s \end{pmatrix}.$$

Furthermore, let $\boldsymbol{\Psi}$ $=cov[(\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}), (\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})]$ so that

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{X}_j\boldsymbol{T}_t & \boldsymbol{X}_j\boldsymbol{T}_{ts} \\ {\boldsymbol{T}'}_{ts}{\boldsymbol{X}'}_j & \boldsymbol{X}_j\boldsymbol{T}_s \end{pmatrix},$$

because $Cov(\boldsymbol{u}_{jt}, \varepsilon_{ijs}) = 0$ by definition. Ψ is a covariance matrix of student proficiencies and random effects for school j. Also, the joint distribution of $\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}, \boldsymbol{u}_{jt}, \boldsymbol{u}_{js}$ is assumed normal with mean $(\boldsymbol{X}_j\boldsymbol{\gamma}_t, \boldsymbol{X}_j\boldsymbol{\gamma}_s, \mathbf{0})$ and covariance matrix

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{M} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi}' & \boldsymbol{G} \end{pmatrix}.$$

Subsequently, the conditional variance of school effects is

$$Var[(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|(\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js})] \quad = \quad \boldsymbol{G} - \boldsymbol{\Psi}'\boldsymbol{M}^{-1}\boldsymbol{\Psi}, \tag{36}$$

using the same theorem as in (23). Furthermore, it can be shown that

$$Cov(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js}) = E\left[Cov(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}\right] - cov\left[E(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}), E(\boldsymbol{u}_{js}|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js})\right]. \tag{37}$$

The covariances in (37), $cov\left[E(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}), E(\boldsymbol{u}_{js}|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js})\right]$, can be evaluated by substituting (23) into the following expression

$$\begin{aligned} &Cov\left[E(\boldsymbol{u}_{jt}|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}), E(\boldsymbol{u}_{js}|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js})\right] = Cov\left[\bar{\boldsymbol{u}}_{jt}, \bar{\boldsymbol{u}}_{js}\right] \\ &= Cov\left[\boldsymbol{C}_{jt}^{-1}{\boldsymbol{X}'}_j(\boldsymbol{\theta}_{jt} - \boldsymbol{X}_j\boldsymbol{\gamma}_t), \boldsymbol{C}_{js}^{-1}{\boldsymbol{X}'}_j(\boldsymbol{\theta}_{js} - \boldsymbol{X}_j\boldsymbol{\gamma}_s)\right] \\ &= \boldsymbol{C}_{jt}^{-1}{\boldsymbol{X}'}_j\left[\sigma_{ts} \otimes \boldsymbol{I}\right]\boldsymbol{X}_j\left[\boldsymbol{C}_{js}^{-1}\right]', \end{aligned} \tag{38}$$

where σ_{ts} are the covariance components of the matrix $\boldsymbol{\Sigma}$. Then, from (36),(37), and (38), the expectation of the conditional covariance $E\left[Cov(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}\right]$ can be expressed as

$$\begin{aligned} E\left[Cov(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}\right] &= \boldsymbol{G} - \boldsymbol{\Psi}' M^{-1} \boldsymbol{\Psi} \\ = Cov(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js}) + \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j \left[\sigma_{ts} \otimes \boldsymbol{I}\right] \boldsymbol{X}_j \left[\boldsymbol{C}_{js}^{-1}\right]' \\ &= \boldsymbol{T}_{st} + \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j \left[\sigma_{ts} \otimes \boldsymbol{I}\right] \boldsymbol{X}_j \left[\boldsymbol{C}_{js}^{-1}\right]'. \end{aligned} \tag{39}$$

Finally, the conditional covariance $Cov\left[(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|\boldsymbol{\Omega}\right]$ is evaluated by

$$\begin{aligned} \tilde{\boldsymbol{T}}_{ts} &= Cov\left[(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|\boldsymbol{\Omega}\right] \\ &= Cov\left[(\boldsymbol{u}_{jt}, \boldsymbol{u}_{js})|\boldsymbol{\theta}_{jt}, \boldsymbol{\theta}_{js}\right] + \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j \left[\tilde{\sigma}_{ij(t,s)} \otimes \boldsymbol{I}\right] \boldsymbol{X}_j \left[\boldsymbol{C}_{jt}^{-1}\right]' \\ &= \boldsymbol{T}_{st} + \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}'_j \left[(\sigma_{ts} + \tilde{\sigma}_{ij(t,s)}) \otimes \boldsymbol{I}\right] \boldsymbol{X}_j \left[\boldsymbol{C}_{js}^{-1}\right]' \end{aligned} \tag{40}$$

with the same method as used in (28). Notice that $\tilde{\sigma}_{ij(t,s)}$ is a posterior covariance component in the posterior covariance matrix $\tilde{\boldsymbol{\Sigma}}_{ij}$, which is given by

$$\tilde{\boldsymbol{\Sigma}}_{ij} = \begin{pmatrix} \tilde{\sigma}_{ij(1,1)} & \tilde{\sigma}_{ij(1,2)} & \tilde{\sigma}_{ij(1,3)} & \cdots & \tilde{\sigma}_{ij(1,p)} \\ \tilde{\sigma}_{ij(2,1)} & \tilde{\sigma}_{ij(2,2)} & \tilde{\sigma}_{ij(2,3)} & \cdots & \tilde{\sigma}_{ij(2,p)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \tilde{\sigma}_{ij(p,1)} & \tilde{\sigma}^2_{ij(p,2)} & \tilde{\sigma}_{ij(p,3)} & \cdots & \tilde{\sigma}_{ij(p,p)} \end{pmatrix}.$$

Hence, to obtain the conditional expectations for T_{st} for the multivariate case, the result from (40) has to be substituted back into (35).

2.2.4 Summary for Multivariate Case

The steps of the EM algorithm for maximum likelihood parameter estimation in the multivariate case can be summarized as follows. The $(r+1)^{th}$ M-step can be completed by computing

$$\boldsymbol{\gamma}_{t,r+1} = \left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij} \boldsymbol{x}'_{ij} \boldsymbol{x}_{ij}\right)^{-1} \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij} \boldsymbol{x}'_{ij} (\tilde{\boldsymbol{\theta}}_{ijt} - \boldsymbol{u}_{jt} \boldsymbol{x}_{ij}), \tag{41}$$

$$\boldsymbol{\Sigma}_{r+1} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij} \tilde{\boldsymbol{\Sigma}}_{ij} + \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij} E[\tilde{\boldsymbol{\varepsilon}}'_{ij} \tilde{\boldsymbol{\varepsilon}}_{ij} | \boldsymbol{\Omega}_r]}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}} \tag{42}$$

and

$$\hat{\boldsymbol{T}}_{tt,r+1} = \frac{1}{J}\sum_{j=1}^{J} \tilde{\boldsymbol{u}}_{jt,r} \tilde{\boldsymbol{u}}'_{jt,r} + \frac{1}{J}\sum_{j=1}^{J} Var(\boldsymbol{u}_{jt,r} | \boldsymbol{\Omega}_r). \tag{43}$$

The E-step can be determined by computing the multivariate posterior moments and, subsequently, $\tilde{\boldsymbol{U}}_j = (\tilde{\boldsymbol{u}}_{j1}, \cdots, \tilde{\boldsymbol{u}}_{jp})$ through (26), $\tilde{\boldsymbol{\varepsilon}}_{ij} = \tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma} - \boldsymbol{x}_{ij}\boldsymbol{U}_j$, $E[\tilde{\boldsymbol{\varepsilon}}'_{ij}\tilde{\boldsymbol{\varepsilon}}_{ij}|\boldsymbol{\Omega}_r]$ through (30), and $Var(\boldsymbol{u}_{jt,r}|\boldsymbol{\Omega}_r)$ through (28). Also, $\tilde{\boldsymbol{u}}_{jt,r}$ and $\boldsymbol{\Omega}_r$ are the conditional mean for $\boldsymbol{u}_{jt}$ and $\boldsymbol{\Omega}$, respectively, at iteration step r.

Equation (43) gives the estimates of p matrices $\hat{\boldsymbol{T}}_{t,r+1}$ at step $r+1$. The other $\frac{p(p-1)}{2}$ covariance matrices of school effects between subscales $t, s = 1, \cdots, p$ are

$$\hat{\boldsymbol{T}}_{ts,r+1} = \frac{1}{J}\sum_{j=1}^{J}\tilde{\boldsymbol{u}}_{jt,r}\tilde{\boldsymbol{u}}'_{js,r} + \frac{1}{J}\sum_{j=1}^{J} Cov[(\boldsymbol{u}_{jt,r}, \boldsymbol{u}_{js,r})|\boldsymbol{\Omega}_r], \tag{44}$$

using (40) to obtain an analytic expression for $Cov[(\boldsymbol{u}_{jt,r}, \boldsymbol{u}_{js,r})|\boldsymbol{\Omega}_r]$. The E- and M-steps are alternated until convergence is obtained. In sum, the general structure of the multivariate case is similar to the univariate case, except for the additional computation of the posterior covariances $\tilde{T}_{st}$ of school effects.

3 Standard Errors

As briefly mentioned in the introduction, programs such as NAEP often make simple random sample assumptions during parameter estimation and then apply a post-hoc complex sample estimator to obtain appropriate standard errors. Hence, initial standard errors of parameters are computed based on simple random sample theory. There are some concerns that deserve further study to improve the estimation of standard errors for regression effects in the current NAEP analysis. Besides ignoring the complex sample structure, affecting both parameter estimates and their standard errors, an approximation is also employed that is governed by assumptions that may not satisfy the complicated NAEP context (e.g., a normal posterior distribution of θ_{ijt}). Before the standard error computation under the HLRM framework is discussed, a brief introduction of the current NAEP method is first provided below. The discussion will be limited to the standard errors of $\boldsymbol{\gamma}$.

3.1 Standard Errors in NAEP

The standard errors of the regression effects are estimated by summing (a) the sampling variance and (b) a variance component that reflects the uncertainty due to the latency of

proficiency. Specifically,

$$\begin{aligned} Var(\hat{\boldsymbol{\gamma}}_t) &\approx Var(\hat{\boldsymbol{\gamma}}_t|\boldsymbol{X},\boldsymbol{Y}_t) \\ &= E[\,Var(\hat{\boldsymbol{\gamma}}_t|\boldsymbol{X},\boldsymbol{Y}_t,\boldsymbol{\theta}_t)] + Var[E(\boldsymbol{\gamma}_t|\boldsymbol{X},\boldsymbol{Y}_t,\boldsymbol{\theta}_t)]. \end{aligned} \tag{45}$$

The first component is estimated assuming that examinees are selected from a simple random sample and examinee proficiency values are observed. For the univariate case, noting that the conditional covariance equals $\sigma_t^2 = E[\varepsilon_i^2] = E[(\tilde{\boldsymbol{\theta}_{it}} - \boldsymbol{\gamma}_t\boldsymbol{x}_i)^2]$, this component is expressed as $(\boldsymbol{X'DX})^{-1}\sigma_t^2$, where $\boldsymbol{D}$ is a diagonal matrix of individual sampling weights. Note that $\tilde{\theta}_{it}$ is the posterior expectation. The second component, taking into account that $\boldsymbol{\theta}_t$ is not observed, is expressed as

$$Var(E(\boldsymbol{\gamma}_t|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}_t)) = (\boldsymbol{X'DX})^{-1}\boldsymbol{X'D}\,Var(\boldsymbol{\theta}_t|\boldsymbol{X},\boldsymbol{Y})\boldsymbol{DX}(\boldsymbol{X'DX})^{-1}, \tag{46}$$

because $E(\boldsymbol{\gamma}_t|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}_t) = (\boldsymbol{X'DX})^{-1}\boldsymbol{X'D}\tilde{\boldsymbol{\theta}}_t$. Hence, the standard errors of regression effects in the univariate case can be approximated by

$$Var(\hat{\boldsymbol{\gamma}}_t) = (\boldsymbol{X'DX})^{-1}\sigma_t^2 + (\boldsymbol{X'DX})^{-1}\boldsymbol{X'D}\tilde{\boldsymbol{\Sigma}}\boldsymbol{DX}(\boldsymbol{X'DX})^{-1}, \tag{47}$$

where $\tilde{\boldsymbol{\Sigma}} = Var(\tilde{\boldsymbol{\theta}}_t|\boldsymbol{X},\boldsymbol{Y})$ is a diagonal matrix with student posterior variances.

For the p-variate case, the covariances between effects across subscales also need to be estimated. The vector of student proficiencies $\boldsymbol{\theta}_i = (\theta_{i1},\cdots,\theta_{ip})'$ for $i = 1,\cdots,N$ is assumed to have a common conditional variance matrix $\tilde{\boldsymbol{\Sigma}}$. The variation due to sampling, $Cov(\hat{\boldsymbol{\gamma}}_s,\hat{\boldsymbol{\gamma}}_t)$, can be expressed as

$$Cov(\hat{\boldsymbol{\gamma}}_s,\hat{\boldsymbol{\gamma}}_t) = E\left[(\hat{\boldsymbol{\gamma}}_s - \boldsymbol{\gamma}_s)(\hat{\boldsymbol{\gamma}}_t - \boldsymbol{\gamma}_t)'\right], \tag{48}$$

and the first component in (45) is computed as

$$Cov(\hat{\boldsymbol{\gamma}}_s,\hat{\boldsymbol{\gamma}}_t) = \sigma_{st}(\boldsymbol{X'DX})^{-1}, \tag{49}$$

where σ_{st} is the (s,t) entry in the covariance matrix

$$\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_{pp} \end{pmatrix}.$$

The second component is similar to the univariate case, except that $\tilde{\boldsymbol{\Sigma}}$ is a collection of $\frac{p(p+1)}{2}$ diagonal matrices.

3.2 Standard Errors for the HLRM

The procedures employed in NAEP can be extended to the hierarchical latent regression model based on (45). The sampling variance (e.g., first component) of the estimate of $\boldsymbol{\gamma}_t$ can be found following a development by Raudenbush and Bryk (2002, p. 42). Assuming $\boldsymbol{X}_j$ to be full column rank, the ordinary least square estimator of $\boldsymbol{\gamma}_{jt}$ is

$$\hat{\boldsymbol{\gamma}}_{jt} \quad = \quad (\boldsymbol{X'}_j\boldsymbol{X}_j)^{-1}\boldsymbol{X'}_j\tilde{\boldsymbol{\theta}}_{jt}, \tag{50}$$

and the dispersion matrix is given by

$$Var(\hat{\boldsymbol{\gamma}}_{jt}) = \boldsymbol{V}_{jt} = (\boldsymbol{X'}_j\boldsymbol{X}_j)^{-1}\sigma_t^2. \tag{51}$$

Multiplying both sides of the Level 1 model in (3) by $(\boldsymbol{X'}_j\boldsymbol{X}_j)^{-1}\boldsymbol{X'}_j$ yields

$$\hat{\boldsymbol{\gamma}}_{jt} = \boldsymbol{\gamma}_{jt} + \boldsymbol{e}_{jt}, \tag{52}$$

where $\boldsymbol{e}_{jt} \sim N(\boldsymbol{0}, \boldsymbol{V}_{jt})$. Moreover, combining the Level 2 model in (4) yields

$$\hat{\boldsymbol{\gamma}}_{jt} = \boldsymbol{\gamma} + \boldsymbol{u}_{jt} + \boldsymbol{e}_{jt}. \tag{53}$$

The variance for $\hat{\boldsymbol{\gamma}}_{jt}$ in (53) is decomposed into two parts: one is the parameter dispersion of $\boldsymbol{u}_{jt}$ (e.g., $\boldsymbol{T}_t$) and the other is the residual dispersion of $\boldsymbol{e}_{jt}$ (e.g., $\boldsymbol{V}_{jt}$). Specifically,

$$Var(\hat{\boldsymbol{\gamma}}_{jt}) = Var(\boldsymbol{u}_{jt} + \boldsymbol{e}_{jt}) = \boldsymbol{T}_t + \boldsymbol{V}_{jt}, \tag{54}$$

is a $Q \times Q$ variance matrix that can be written as

$$Var(\hat{\boldsymbol{\gamma}}_{jt}) = \triangle_{jt} = \boldsymbol{T}_t + (\boldsymbol{X'}_j\boldsymbol{X}_j)^{-1}\sigma_t^2. \tag{55}$$

In most educational surveys, school or cluster sample sizes are not balanced, and the values for $\triangle_{jt}$ will differ from school to school. Assuming $\triangle_{jt}$ is known, the unique, minimum variance, unbiased estimator of $\boldsymbol{\gamma}_t$ will be the generalized least square estimator

$$\hat{\boldsymbol{\gamma}}_t \quad = \quad (\sum_{j=1}^{J} \triangle_{jt}^{-1})^{-1} \sum_{j=1}^{J} \triangle_{jt}^{-1}\hat{\boldsymbol{\gamma}}_{jt}. \tag{56}$$

Subsequently, the variance of $\hat{\boldsymbol{\gamma}}_t$ is

$$Var(\hat{\boldsymbol{\gamma}}_t) = (\sum_{j=1}^{J} \triangle_{jt}^{-1})^{-1}. \tag{57}$$

If $\triangle_{jt}$ is not known, it can be estimated by (55) using $\tilde{\boldsymbol{T}}$ for $\boldsymbol{T}$.

The second component in (45) addresses the variation due to the latency of $\boldsymbol{\theta}$ and is equal to the variance of the posterior expectation of $\hat{\boldsymbol{\gamma}}_t$. The posterior expectation of $\hat{\boldsymbol{\gamma}}_t$ is equivalent to

$$E(\boldsymbol{\gamma}_t|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}) \;=\; (\sum_{j=1}^{J}\triangle_{jt}^{-1})^{-1}\sum_{j=1}^{J}[\triangle_{jt}^{-1}(\boldsymbol{X'}_j\boldsymbol{X}_j)^{-1}\boldsymbol{X'}_j\boldsymbol{\theta}_j]. \tag{58}$$

Subsequently, the variation of this expectation can be estimated by

$$Var(E(\hat{\boldsymbol{\gamma}}_t|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta})) = \boldsymbol{V}\sum_{j=1}^{J}\left[\triangle_j^{-1}(\boldsymbol{X'}_j\boldsymbol{X}_j)^{-1}\boldsymbol{X'}_j\tilde{\boldsymbol{\Sigma}}_j\boldsymbol{X}_j(\boldsymbol{X'}_j\boldsymbol{X}_j)^{-1}\triangle_j^{-1'}\right]\boldsymbol{V'}, \tag{59}$$

for $\boldsymbol{V} = [(\sum_{j=1}^{J}\triangle_j^{-1})]^{-1}$.

It should be noted that the formula for computing $E(\,Var(\hat{\boldsymbol{\gamma}}_t|\boldsymbol{X},\boldsymbol{Y}_t,\boldsymbol{\theta}_t))$ in (57) and $Var(E(\hat{\boldsymbol{\gamma}}_t|\boldsymbol{X},\boldsymbol{Y}_t,\boldsymbol{\theta}_t))$ in (59) are feasible if all Level 1 coefficients are considered random and every Level 2 unit contains an adequate sample to computate $\hat{\boldsymbol{\gamma}}_{jt}$ in (50). Raudenbush and Bryk (2002, pp. 44–45) stated that the formulae for computing standard errors for $\hat{\boldsymbol{\gamma}}_t$ have rather limited use in practical applications. Following their derivations (pp. 44–45), the generalized least squares estimator for fixed effects is

$$\hat{\boldsymbol{\gamma}}_t = \left(\sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{V}_{\boldsymbol{\theta}_j}^{-1}\boldsymbol{X}_j\right)^{-1}\sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{V}_{\boldsymbol{\theta}_j}^{-1}\boldsymbol{X}_j, \tag{60}$$

where

$$V_{\boldsymbol{\theta}_j} = Var(\boldsymbol{\theta}_j) = \boldsymbol{X}_j\boldsymbol{T}\boldsymbol{X}'_j + \sigma^2\boldsymbol{I}. \tag{61}$$

Hence, the variance covariance matrix of for fixed effects estimates $\hat{\boldsymbol{\gamma}}_t$ is

$$Var(\hat{\boldsymbol{\gamma}}_t) = \left(\sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{V}_{\boldsymbol{\theta}_j}^{-1}\boldsymbol{X}_j\right)^{-1}. \tag{62}$$

If $\boldsymbol{X}_j$ is full rank, the result in (57) is equivalent to the result from (62), as proven by Raudenbush and Bryk (2002, p. 44).

4 Application to NAEP Data

The NAEP 2004 age 17 long-term trend mathematics assessment data has been analyzed with a two-level simple HLRM model without predictors on the second level. The data contains 7,561 students, and a single scale is assumed under the IRT model employed here. There are 62

cluster units, with sample sizes for each unit ranging from 3 students to 291 students. There are 7 clusters where the number of students is less than 30, and 5 clusters where it is less than 20. The regression parameter estimates, as stated before, are overall mean effects across clusters. Thus, the estimates of regression effects parameters are expected to be close to the estimates from the current NAEP approach.

A large model and three small models are tested. The large model contains 156 independent predictors plus an intercept at the first-level model. The predictors are represented as principal components x^* with standard deviations sd^*, where $sd_1^* \geq sd_2^* \geq \cdots \geq sd_q^* \geq \cdots \geq sd_Q^*$. These principal components are extracted from dummy codes indicating membership to many student groups. The three small models contain only student group membership indicators or contrasts for (a) gender (male vs. female) and (b) racial ethnicity (White, Black, Hispanic, Asian/others), and (c) gender + race/ethnicity, in addition to the intercept included in each model.

An EM algorithm for parameter estimation of a simple HLRM model is implemented via a C++ program.

4.1 Small Models

Small Model 1: Gender. A small model was estimated, using students' gender designations as predictors at the first level, modeling male versus female students. The residual variance estimate in this small model is .9145, obliviously greater than that of the large operational model containing 157 principal components, which is .3737. The larger residual variance of this small model implies that a large amount of the variation of latent traits is not accounted by this indicator. The residual variance estimate under the current NAEP approach is .978, which is slightly larger than that from the HLRM approach, implying that the variation accounted for by the clusters is small with respect to the male/female distinction.

The regression effect and intercept estimates and standard errors, as well as the estimates from the current NAEP approach, are given in Table 1. The intercept estimate is .075, and regression effect for female students is -.1104. The estimates of regression effects are very close to those from current NAEP approach, with slight differences appearing in the fourth decimal place. Both regression effect estimates (HLM and NAEP) for female are negative, implying male students generally perform better than female students.

Columns 4 and 5 are standard errors estimates for the regression parameters, denoted by

Table 1
HLM Parameter Estimates Versus NAEP (Gender)

Variable	NAEP($\hat{\gamma}$)	HLRM($\hat{\gamma}$)	HLRM SE($\hat{\gamma}$)	HLRM SE_1($\hat{\gamma}$)
Intercept	.075	.0752	.0498	.0496
Female	-.1104	-.1105	.0232	.0226

$HLRMSE(\hat{\gamma})$ and $HLRMSE_1(\hat{\gamma})$, respectively. $HLRMSE_1(\hat{\gamma})$ involves sampling variation only, but $HLRMSE(\hat{\gamma})$ incorporates both sampling and measurement variation. It shows from the table that almost all the variation for estimating the regression parameters is attributed to sampling; only a small amount of variation is due to the latency of a student ability.

Small Model 2: Race/ethnicity. A small model was also estimated using students' racial group designations as predictors at the first level, modeling White students versus other races of students (Black, Hispanic, Asian/others). The residual variance estimate in this small model is .8312, obliviously greater than that of the large operational model containing 157 principal components, which is .3737. The larger residual variance of this small model implies that a large amount of the variation of latent traits is not accounted for by indicators of race/ethnicity. The residual variance estimate under the current NAEP approach is .8674, which is slightly larger than that from the HLRM approach, implying that the variation accounted for by the clusters is small with respect to the racial distinctions.

The regression effect, intercept estimates, and standard errors, as well as the estimates from the current NAEP approach, are given in Table 2. The intercept estimate is .208 and regression effect for Black students is -.8688; for Hispanic students, -.6306; and for Asian/other students, .0691. The estimates of regression effects are very close to those from the current NAEP approach, with slight differences appearing in the fourth decimal place. Both regression effect estimates (HLRM and NAEP) for Black and Hispanic students are negative, implying White students generally perform better than Black and Hispanic students.

Small Model 3: Gender + race/ethnicity. A small model was also estimated using both students' gender and racial ethnicity designations as predictors at the first level, modeling male versus female and White versus other racial students (Black, Hispanic, Asian/others). The residual variance estimate in this small model is .8290, obliviously greater than that of the large operational model containing 157 principal components, which is .3737. The larger residual

Table 2
HLM Parameter Estimates Versus NAEP (Racial)

Variable	NAEP($\hat{\gamma}$)	HLRM($\hat{\gamma}$)	HLRM SE($\hat{\gamma}$)	HLRM $SE_1(\hat{\gamma})$
Intercept	.208	.2083	.0358	.0356
Black	-.868	-.8688	.0441	.0433
Hispanic	-.630	-.6306	.0496	.0488
Asian/others	.069	.0691	.0665	.0652

variance of this small model implies that a large amount of the variation of latent traits is not accounted for by indicators of racial ethnicity and gender. The residual variance estimate under the current NAEP approach is .8653, which is slightly larger than that from the HLRM approach, implying that the variation accounted for by the clusters is small with respect to the gender and racial distinctions.

The regression effect, intercept estimates, and standard errors as well as the estimates from the current NAEP approach are given in Table 3. The intercept estimate is .2514. The regression effect for female students is -.0882; for Black students, -.8641; for Hispanic students, -.6279, and for Asian/other students, .0729. The estimates of regression effects are very close to those from the current NAEP approach, with slight differences appearing in the fourth decimal place. Both regression effect estimates (HLRM and NAEP) for female students are negative, implying male students generally perform better than female students. Both regression effects estimates (HLRM and NAEP) for Black and Hispanic students are negative, implying White students generally perform better than Black and Hispanic students.

Table 3
HLM Parameter Estimates Versus NAEP (Gender + Racial)

Variable	NAEP($\hat{\gamma}$)	HLRM($\hat{\gamma}$)	HLRM SE($\hat{\gamma}$)	HLRM $SE_1(\hat{\gamma})$
Intercept	.2514	.2516	.03959	.0393
Female	-.0882	-.0882	.0225	.0219
Black	-.864	-.8641	.0432	.0424
Hispanic	-.627	-.6279	.0491	.0483
Asian/others	.0729	.0729	.0668	.0656

4.2 Large Model

Model parameter estimates. A large model was estimated using 156 principal components extracted from the covariance matrix of the student group indicators employed in the current NAEP model. This model converged around the 100th EM cycle, with the total likelihood monotonically increasing until convergence, as is presented in Figure 1. The HLRM regression parameter estimates plus the intercept for the large model are given in Table 4, Table 5, Table 6, and Table 7 in column 5, denoted $HLM(\hat{\gamma})$, along with current NAEP estimates of regression parameters in the second column, denoted as $NAEP(\hat{\gamma})$. (Note that Table 4, Table 5, Table 6, and Table 7 actually constitute a long table, containing the results of parameter estimates with a large set of principle components extracted from operational NAEP analysis). The regression parameter estimates for these two approaches are very close to each other, which is expected, since the HLM regression effects are the overall mean estimates of the regression effects for each cluster. Figure 2 displays a plot of the regression effects estimates from HLRM versus those from NAEP.

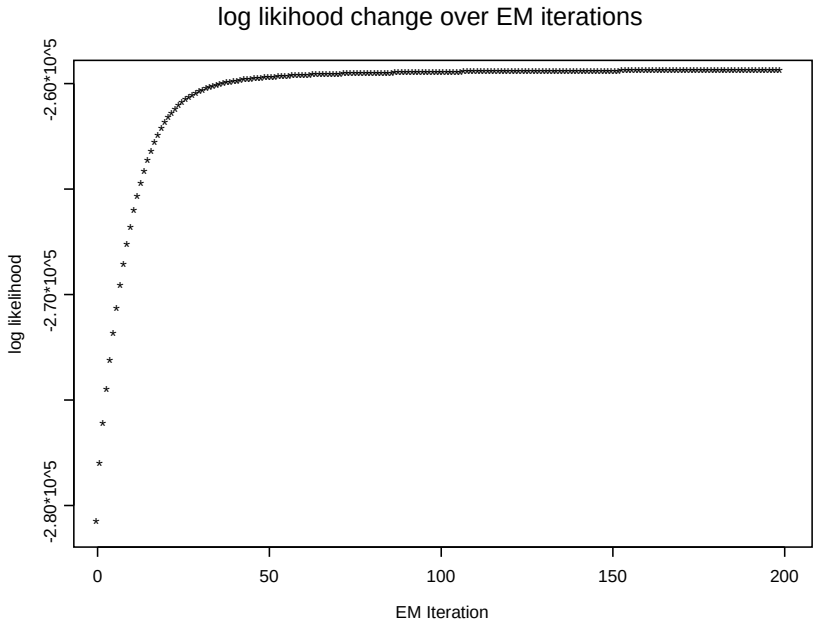

***Figure* 1 The log likelihood for the first 200 EM iterations.**

The estimate of the residual variance from the simple HLRM model is .3737, compared to .567 in the current NAEP model, which is expected because the HLRM takes the variation across

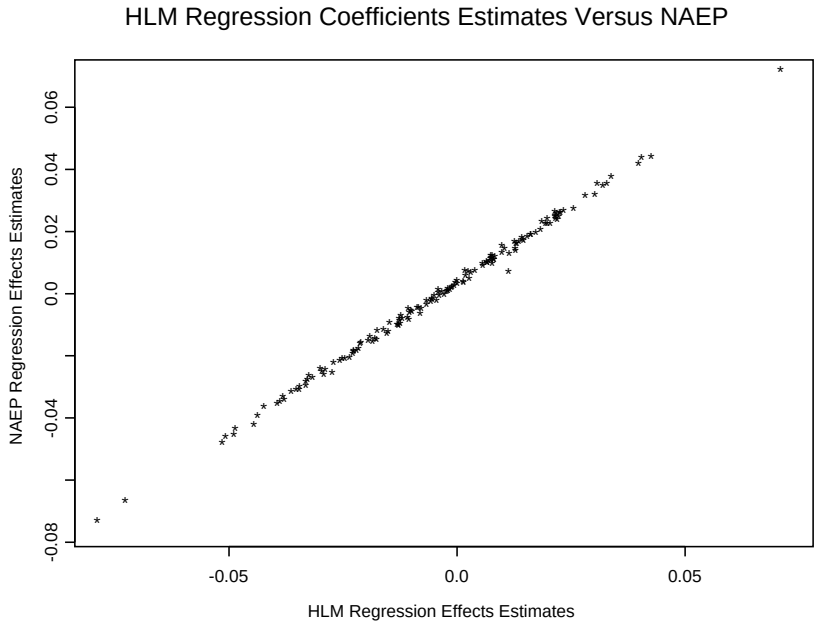

***Figure* 2 Comparison of HLM estimates for regression effects with NAEP.**

clusters into account. Not entirely expected is the formidable magnitude of this difference. The comparison of the residual variance estimates among the large model and the three small models is given in Table 8.

Although the covariance matrix among random effects $\boldsymbol{T}_t$ has dimension 157, the computation of $\hat{\boldsymbol{T}}_t$ requires only the first two posterior moments of $\boldsymbol{u}$. Therefore, the computations of $\hat{\boldsymbol{T}}$ is stable as long as the computation of $\tilde{\boldsymbol{u}}_{jt}$ and $Var(\boldsymbol{u}_{jt}|\boldsymbol{\Omega})$ are stable. Because $\hat{\boldsymbol{T}}_t$ is a large matrix, the elements are not listed here. However, the diagonal components range from .000069 to .0142, which implies that there exist substantial differences between clusters, supporting the hierarchical model. It seems prudent to establish a mechanism to determine the significance of incorporating random effects in a given analysis.

Standard errors. The standard error estimates from the simple two-level HLRM are expected to be higher than the current NAEP estimates because additional variation across clusters are accounted for. In Tables 4 through 7, column 3 denotes the NAEP standard error $NAEPSE(\hat{\gamma})$, and column 6 denotes the HLRM standard error, $HLMSE(\hat{\gamma})$. Furthermore, in Tables 4 through 7, columns 4 and 7 show the standard error estimates for the part due to sampling only. From

Table 4

HLRM Parameter Estimates Versus NAEP Estimates

PCF	$NAEP(\hat{\gamma})$	$NAEPSE(\hat{\gamma})$	$NAEPSE_1(\hat{\gamma})$	$HLM(\hat{\gamma})$	$HLMSE(\hat{\gamma})$	$HLMSE_1(\hat{\gamma})$
1	.0046	.009	.0087	.0122	.0219	.0211
2	-.0374	.001	.0009	-.0377	.0022	.0021
3	.0049	.0013	.0012	.005	.0032	.0031
4	.0142	.0016	.0015	.0137	.0034	.0033
5	-.0171	.0018	.0018	-.0169	.0035	.0034
6	-.0461	.0019	.0018	-.0474	.0046	.0045
7	-.0017	.0019	.0019	-.0012	.0024	.0023
8	-.0235	.0021	.002	-.0241	.005	.0049
9	-.001	.0021	.0021	-.0009	.0032	.0031
10	.0013	.0023	.0022	.0024	.0039	.0037
11	.0223	.0023	.0022	.0226	.0041	.0039
12	.009	.0023	.0022	.0085	.0036	.0035
13	-.0091	.0025	.0023	-.007	.0052	.0049
14	.0149	.0025	.0024	.0154	.0042	.0041
15	-.0071	.0027	.0026	-.0076	.0065	.0064
16	.0158	.0027	.0026	.0165	.0034	.0033
17	-.0011	.0027	.0026	-.0005	.0051	.0049
18	-.0692	.0028	.0027	-.0716	.0044	.0043
19	.0234	.0028	.0027	.0236	.0052	.005
20	.0075	.0029	.0028	.0077	.0052	.005
21	-.0207	.0029	.0028	-.021	.0062	.006
22	-.0084	.003	.0029	-.0088	.0055	.0053
23	-.0341	.0031	.003	-.0353	.006	.0058
24	-.0122	.0032	.0031	-.0113	.0056	.0054
25	.001	.0033	.0032	.0007	.0132	.0129
26	.0032	.0034	.0033	.003	.0063	.0061
27	.008	.0035	.0034	.0085	.0056	.0054
28	.0093	.0038	.0036	.0094	.0078	.0076
29	-.0324	.0039	.0037	-.0334	.0069	.0066
30	.0085	.0039	.0038	.0092	.0067	.0064
31	-.0241	.004	.0039	-.0246	.0075	.0073
32	.0141	.0042	.004	.0147	.0074	.0072
33	-.012	.0043	.0041	-.0137	.0081	.0078
34	-.0285	.0044	.0042	-.0281	.009	.0088
35	-.029	.0044	.0042	-.0314	.0076	.0073
36	.0392	.0044	.0043	.0409	.0089	.0085
37	.0083	.0045	.0043	.0092	.0093	.009
38	.0043	.0045	.0044	.0041	.0081	.0078
39	-.0233	.0047	.0045	-.0235	.0085	.0082
40	.0416	.0047	.0045	.0436	.0081	.0078
41	-.0378	.0047	.0046	-.0383	.0091	.0088
42	.0164	.0049	.0047	.0173	.009	.0086
43	.0181	.0049	.0048	.0194	.0087	.0084
44	.0222	.005	.0048	.0234	.0086	.0083
45	-.027	.005	.0048	-.0278	.0078	.0075
46	.0155	.0051	.0049	.0153	.008	.0077
47	.0171	.0051	.0049	.0183	.01	.0096

Table 5

HLRM Parameter Estimates Versus NAEP Estimates–Continued

PCF	$NAEP(\hat{\gamma})$	$NAEPSE(\hat{\gamma})$	$NAEPSE_1(\hat{\gamma})$	$HLM(\hat{\gamma})$	$HLMSE(\hat{\gamma})$	$HLMSE_1(\hat{\gamma})$
48	-.0367	.0051	.005	-.0367	.0091	.0088
49	-.0059	.0052	.005	-.0056	.0091	.0088
50	-.0015	.0053	.0051	-.001	.0082	.0078
51	-.0276	.0053	.0052	-.0286	.0086	.0083
52	-.0027	.0054	.0052	-.0017	.0116	.0113
53	-.0043	.0054	.0052	-.0044	.0093	.009
54	.02	.0055	.0053	.0216	.0094	.0091
55	-.0113	.0056	.0054	-.0115	.0103	.01
56	-.0106	.0056	.0054	-.0109	.0088	.0085
57	-.0212	.0057	.0055	-.0216	.0096	.0093
58	-.0069	.0057	.0055	-.0074	.0085	.0082
59	.0088	.0058	.0056	.0084	.0076	.0073
60	-.0015	.0058	.0056	-.001	.0103	.0098
61	-.0175	.0059	.0057	-.0184	.0082	.0079
62	.0214	.0059	.0057	.023	.01	.0096
63	-.002	.006	.0058	-.0022	.0104	.01
64	.0137	.0061	.0059	.014	.0126	.0122
65	.0144	.0062	.006	.0157	.0089	.0085
66	.0107	.0062	.006	.0109	.0123	.0119
67	-.0048	.0063	.0061	-.0034	.0117	.0113
68	-.0265	.0064	.0061	-.0289	.0113	.0109
69	-.0022	.0064	.0062	-.003	.0126	.0122
70	-.0144	.0064	.0062	-.0164	.0126	.0123
71	.0072	.0065	.0063	.0088	.0107	.0104
72	-.0003	.0065	.0063	0	.0094	.009
73	-.0505	.0066	.0063	-.0504	.0122	.0117
74	.0232	.0066	.0064	.0229	.0122	.0118
75	-.0756	.0067	.0064	-.0777	.0094	.009
76	-.032	.0068	.0065	-.032	.0116	.0112
77	-.0246	.0068	.0066	-.0259	.0112	.0107
78	.0023	.0069	.0066	.0038	.0148	.0145
79	.001	.0069	.0067	.0025	.0108	.0103
80	.0694	.007	.0068	.072	.0102	.0097
81	.0073	.0071	.0068	.0067	.0126	.0121
82	-.0445	.0071	.0068	-.0435	.0135	.0131
83	.0064	.0071	.0069	.0068	.0138	.0133
84	.0118	.0072	.0069	.0115	.0124	.0119
85	.0215	.0072	.007	.0209	.0139	.0134
86	.0292	.0073	.007	.0312	.0114	.0109
87	.0207	.0073	.0071	.0197	.014	.0136
88	-.031	.0074	.0071	-.0321	.0139	.0135
89	.024	.0075	.0072	.0225	.011	.0106
90	-.0477	.0075	.0072	-.0478	.0133	.0129
91	.0016	.0075	.0072	.001	.0123	.0119
92	-.0203	.0075	.0073	-.0205	.0116	.0112
93	-.0128	.0076	.0073	-.0117	.0105	.01
94	.0329	.0076	.0073	.0339	.0132	.0127

Table 6

HLRM Parameter Estimates Versus NAEP Estimates—Continued

PCF	$NAEP(\hat{\gamma})$	$NAEPSE(\hat{\gamma})$	$NAEPSE_1(\hat{\gamma})$	$HLM(\hat{\gamma})$	$HLMSE(\hat{\gamma})$	$HLMSE_1(\hat{\gamma})$
95	.0216	.0077	.0075	.0227	.0135	.013
96	.0324	.0078	.0075	.033	.0138	.0134
97	-.0108	.0078	.0075	-.0095	.0129	.0124
98	.0199	.0078	.0075	.0209	.0116	.0112
99	-.0388	.0078	.0075	-.0412	.0113	.0108
100	.029	.0079	.0076	.0292	.013	.0125
101	.0241	.008	.0077	.0245	.0143	.0138
102	-.0048	.008	.0077	-.0056	.0128	.0124
103	-.014	.0081	.0078	-.015	.0136	.0131
104	-.0031	.0081	.0078	-.0039	.0121	.0116
105	-.004	.0082	.0079	-.0044	.0151	.0147
106	-.0148	.0082	.0079	-.014	.0118	.0113
107	-.023	.0082	.0079	-.0225	.0128	.0124
108	-.0107	.0083	.0079	-.0116	.0123	.0117
109	.0008	.0083	.008	.0011	.0133	.0128
110	.0237	.0083	.008	.0238	.0148	.0142
111	.0137	.0083	.008	.0142	.0149	.0144
112	-.0101	.0083	.008	-.0098	.016	.0156
113	-.0163	.0084	.0081	-.018	.0139	.0134
114	-.0419	.0084	.0081	-.0426	.0141	.0136
115	.0104	.0084	.0081	.0125	.0144	.014
116	-.0303	.0085	.0082	-.0316	.0153	.0149
117	.0227	.0085	.0082	.0226	.0151	.0146
118	.0114	.0086	.0083	.0139	.0168	.0163
119	-.0174	.0086	.0083	-.0165	.0142	.0137
120	-.0335	.0086	.0083	-.0341	.0153	.0147
121	-.0079	.0087	.0084	-.009	.0183	.0177
122	-.0052	.0087	.0084	-.0046	.0164	.0159
123	-.0297	.0088	.0084	-.0306	.0172	.0167
124	-.0484	.0088	.0085	-.0496	.015	.0144
125	-.0125	.0088	.0085	-.0118	.0182	.0176
126	-.0356	.0089	.0086	-.037	.0138	.0132
127	.0248	.0089	.0086	.0266	.0123	.0117
128	-.0075	.009	.0086	-.0067	.0144	.0139
129	.0199	.009	.0087	.0205	.0138	.0132
130	.0045	.0091	.0087	.0035	.0147	.0141
131	.0121	.0091	.0087	.0139	.0129	.0123
132	-.0072	.0091	.0088	-.0097	.0186	.018
133	.0352	.0091	.0088	.0349	.0147	.0141
134	.0412	.0092	.0088	.0416	.0207	.0202
135	.0129	.0092	.0089	.011	.0149	.0145
136	.0049	.0093	.0089	.0028	.0159	.0153
137	-.0124	.0093	.009	-.012	.017	.0163
138	-.0187	.0093	.009	-.0201	.0202	.0198
139	-.0333	.0093	.009	-.0335	.0152	.0147
140	-.0086	.0094	.009	-.0093	.0155	.0149
141	-.0095	.0094	.0091	-.0112	.0165	.0159

Table 7
HLRM Parameter Estimates Versus NAEP Estimates—Continued

PCF	$NAEP(\hat{\gamma})$	$NAEPSE(\hat{\gamma})$	$NAEPSE_1(\hat{\gamma})$	$HLM(\hat{\gamma})$	$HLMSE(\hat{\gamma})$	$HLMSE_1(\hat{\gamma})$
142	.0098	.0095	.0091	.0088	.0142	.0136
143	-.0208	.0095	.0091	-.0216	.0149	.0144
144	.0166	.0095	.0092	.0172	.0169	.0164
145	.033	.0096	.0093	.0319	.017	.0164
146	.0078	.0096	.0093	.0076	.0135	.013
147	-.0007	.0097	.0093	-.0001	.0137	.0132
148	-.0014	.0097	.0094	-.003	.0157	.0151
149	-.0179	.0098	.0094	-.0174	.0174	.0169
150	-.0217	.0098	.0095	-.0218	.0154	.0149
151	-.0032	.0099	.0096	-.0029	.0155	.015
152	.0074	.0099	.0096	.0077	.0132	.0126
153	.0002	.0101	.0097	.0004	.0154	.0149
154	.0098	.0101	.0097	.0085	.0147	.0142
155	-.0184	.0101	.0097	-.02	.0157	.0152
156	-.028	.0101	.0098	-.0263	.0177	.0171
157	-.0155	.0102	.0098	-.0143	.0145	.0138

Table 8
HLM Residual Variance Estimates $\hat{\sigma}^2$ Versus NAEP

$Model$	Large	Gender	Racial	Gender + Racial
HLM	.3737	.9146	.8313	.8290
NAEP	.5673	978	.8674	.8653

these columns it can be seen that the HLRM standard errors are substantially larger than standard errors from the current NAEP approach, falling a range between 2 and 10 times as large. Hence, as expected the current NAEP standard error estimates for regression effects are underestimates.

As stated in Section 3, the variance of $\hat{\gamma}$ contains both variation due to sampling and variation due to measurement errors for student abilities. For NAEP estimates for standard errors, the variation due to sampling is much larger than the variation due to the latency, where the former typically accounts for about 90% to 95% of the estimates. Similar pattern of results are found with the HLRM model standard error estimates, although the proportion due to sampling is higher. In Tables 4 through 7, the variation due to sampling is specifically listed as $HLMSE_1(\hat{\gamma})$ for the HLRM model and $NAEPSE_1(\hat{\gamma})$ for the NAEP model. The proportion due to sampling in the HLRM model ranges from 94.73% to 97.83%, slightly higher than the proportions within the current NAEP approach, partly due to the variation across clusters being accounted for and added to the sampling variation.

5 Discussion and Conclusion

In this paper a hierarchical latent regression model has been developed for use in large-scale assessments such as NAEP. The primary purpose of this model is to account for the hierarchical nature of the sample, hence, improving regression parameters and standard errors estimates.

A simple two-level HLRM model discussed in this paper can be easily extended to more general two-level and/or higher level hierarchical linear models incorporating IRT modeling for student latent abilities, as is the case for the current NAEP models. The HLRM is naturally adapted from the current NAEP model, as students are naturally nested within schools.

The regression effect estimates from a simple two-level HLRM model can be compared directly with the current NAEP estimates. However, they are by design more appropriate, since school clustering is taken into account. Some indication of this is provided by a crude comparison of standard errors, which appear under the HLRM to be at least twice the size of the current NAEP estimates. Another indicator is the fact that the residual variance decreased in the simple HLRM model (e.g., the three small models and the large model discussed above) compared to the current NAEP estimates, which is expected, since the random effects term across clusters accounts for variation that is otherwise attributed to unexplained variation.

However, the interest was not limited to parameter estimates, but also included the general feasibility of estimating these parameters. Under the proposed formulation of the HLRM, the $\boldsymbol{T}$ matrix estimate provides some indication of the virtues of employing a hierarchical model. First of all, in the applications presented in this study, the diagonal of this matrix was substantial, indicating a nontrivial random effect. Second, no specific problems were encountered during estimation. Specifically, the estimation of $\hat{\boldsymbol{T}}$ involves only computing $\tilde{\boldsymbol{u}}_{jt}$, and $Var(\tilde{\boldsymbol{u}_{jt}}|\boldsymbol{\Omega})$, the posterior mean and variance of $\boldsymbol{u}_{jt}$ and can yield a numerically stable estimate as long as those moments can be computed in a preceding stage.

With the same convergence criterion, the EM algorithm in the current NAEP procedure takes six iterations to reach convergence, while the algorithm for HLM estimation needs many more cycles to reach convergence (more than 50 in this example) for the large model discussed. Figure 1 is a plot of likelihood changes over the first 200 EM iterations. It shows that the *log* likelihood function is monotonically increasing over the first 200 iterations, which implies that the likelihood increases monotonically cycle by cycle as well. One possible explanation for this slow convergence could be the insufficient number of students in some clusters. As is addressed before, we have 62

clusters in this example, but 5 cluster units have less than 20 students and 7 cluster units have less than 30 students.

6 Future Research Directions

The purpose of regression parameter estimates and standard error estimates is to provide reasonable estimates for NAEP target statistics for population characteristics (e.g., mean scale scores for subgroups, percentage above a certain level of performance achievement). How the parameter estimation in the simple HLRM model would affect NAEP reporting scale scores and their standard errors for each subgroup is of great interest and currently under study.

An important assumption of the HLRM model is that the item parameters are assumed to be fixed or estimated without errors, which is obviously an unsatisfactory statement. Simultaneous estimation of IRT item parameters and HLRM regression parameters also constitutes work in progress.

References

Adams, R. J., Wilson, M. R., & Wu, M. L. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 46–75.

Aitkin, M., & Aitkin, I. (2005). *Comparison of direct estimation with the conditioning model and plausible value imputation.* Paper prepared for U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (No. NCES 2001-509). Washington, DC: national Center for Educational Statistics.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411–434.

Birnbaum, A. R. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Johnson, M. S. (2002, April). *A Bayesian hierarchical model for multidimensional performance assessments.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Johnson, M. S., & Jenkins, F. (2005). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress* (ETS Research Rep. No. RR-04-38). Princeton, NJ: ETS.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993–997.

Raudenbush, S. W., & Byk, A. S. (2002). *Hierarchical linear models: Applications and data analysis method* (2nd ed.). New Delhi, India: Sage Publications India Pvt Ltd.

Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1999). Synthesizing results from the trial state assessment. *Journal of Educational and Behavioral Statistics, 24*(4), 413–438.

Stapleton, J. H. (1995). *Linear statistical models.* New York: John Wiley & Sons, Inc.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wilson, M., & Adams, R. (1995). Issues in complex sampling involving latent variables. In *Proceedings of the Survey Research Methods Section, American Statistical Association,* 263–268.

Appendix

In the multivariate case, regression effects $\mathbf{\Gamma}$ will become a matrix including the regression effects for each subscale (i.e., $\mathbf{\Gamma}$ is a $Q \times p$ matrix $[\boldsymbol{\gamma}_1|, \cdots, |\boldsymbol{\gamma}_p]$ with each subscale regression effects $\boldsymbol{\gamma}_t$ having Q components for $t = 1, \cdots, p$). Let $\boldsymbol{x}_{ij}$ be the collection (or a row) of background variables for student i in school j in the p-scale assessment, then the likelihood function L for N students' responses to n items in the test given school random effects $\boldsymbol{U}_j$ is the total marginal likelihood, and is expressed as

$$\begin{aligned} L &= log\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} P(\boldsymbol{y}_{ij}|\mathbf{\Gamma'}_{ij}\boldsymbol{x}'_{ij} + \boldsymbol{U}'_j\boldsymbol{x}'_{ij}, \mathbf{\Sigma})^{w_{ij}}\right] \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij} log\left[P(\boldsymbol{y}_{ij}|\mathbf{\Gamma'}_{ij}\boldsymbol{x}'_{ij} + \boldsymbol{U}'_j\boldsymbol{x}'_{ij}, \mathbf{\Sigma})\right] \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij} log\left[\int P(\boldsymbol{y}_{ij}|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\mathbf{\Gamma'}_{ij}\boldsymbol{x}'_{ij} + \boldsymbol{U}'_j\boldsymbol{x}'_{ij}, \mathbf{\Sigma})d\boldsymbol{\theta}\right]. \end{aligned} \tag{63}$$

$\phi(\boldsymbol{\theta}|\mathbf{\Gamma'}_{ij}\boldsymbol{x}'_{ij} + \boldsymbol{U}'_j\boldsymbol{x}'_{ij}, \mathbf{\Sigma})$ represents for the conditional multivariate normal density with mean vector $\mathbf{\Gamma'}_{ij}\boldsymbol{x}'_{ij} + \boldsymbol{U}'_j\boldsymbol{x}'_{ij}$ and covariance matrix $\mathbf{\Sigma}$ (i.e., $\boldsymbol{\theta}|\boldsymbol{U}_j \sim$
$\mathcal{N}(\mathbf{\Gamma'}_{ij}\boldsymbol{x}'_{ij} + \boldsymbol{U}'_j\boldsymbol{x}'_{ij}, \mathbf{\Sigma})$). Denote the expectation of $\boldsymbol{\theta}$ by $\boldsymbol{\mu}_\theta = \mathbf{\Gamma'}_{ij}\boldsymbol{x}'_{ij} + \boldsymbol{U}'_j\boldsymbol{x}'_{ij}$, then the density function is given by

$$\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)'\mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)\right]. \tag{64}$$

The partial derivative of $log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \mathbf{\Sigma})$ with respect to $\mathbf{\Gamma'}$ is

$$\frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \mathbf{\Sigma})}{\partial \mathbf{\Gamma'}} = \mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \mathbf{\Gamma'}\boldsymbol{x}'_{ij})\boldsymbol{x}'_{ij}. \tag{65}$$

Therefore,

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{\Gamma'}} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\int \frac{P(\boldsymbol{y}_{ij}|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \mathbf{\Sigma})}{P(\boldsymbol{y}_{ij})}\frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \mathbf{\Sigma})}{\partial \mathbf{\Gamma'}}d\boldsymbol{\theta} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\int \frac{P(\boldsymbol{y}_{ij}|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \mathbf{\Sigma})}{P(\boldsymbol{y}_{ij})}\mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)\boldsymbol{x}_{ij}d\boldsymbol{\theta} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\int P(\boldsymbol{\theta}|\boldsymbol{y}_{ij})\mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)\boldsymbol{x}_{ij}d\boldsymbol{\theta} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\mathbf{\Sigma}^{-1}(\tilde{\boldsymbol{\theta}} - \mathbf{\Gamma'}\boldsymbol{x}'_{ij} - \boldsymbol{U}_j\boldsymbol{x}'_{ij})\boldsymbol{x}_{ij}. \end{aligned} \tag{66}$$

Set (57) to 0, and then we can obtain the estimates of each subscale $\boldsymbol{\gamma}_t$, $t = 1, \cdots, p$, as is given by

$$\hat{\boldsymbol{\gamma}}_t = \left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \boldsymbol{x}'_{ij} \boldsymbol{x}_{ij} \right)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \boldsymbol{x}'_{ij} (\tilde{\theta}_{ijt} - \boldsymbol{u}_{jt} \boldsymbol{x}'_{ij}). \tag{67}$$

To obtain the estimates of $\boldsymbol{\Sigma}$, it follows that

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\Sigma}} &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \int \frac{P(\boldsymbol{y}_{ij}|\boldsymbol{\theta}) \phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)}{P(\boldsymbol{y}_{ij})} \frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)}{\partial \boldsymbol{\Sigma}} d\boldsymbol{\theta} \\ &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \int p(\boldsymbol{\theta}|\boldsymbol{y}_{ij}) \frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} d\boldsymbol{\theta}. \end{aligned} \tag{68}$$

In the multivariate case, $\boldsymbol{U}_j$, like $\boldsymbol{\Gamma}$, is a matrix formed by columns of school random effects for each subscale (i.e., $\boldsymbol{U}_j = [\boldsymbol{u}_{1j}|, \boldsymbol{u}_{2j}|, \cdots |, \boldsymbol{u}_{pj}]$). Now it becomes more convenient to find the derivatives of $log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$, a symmetric matrix. Denote the matrix $\boldsymbol{\Xi} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)'\boldsymbol{\Sigma}^{-1}$, then the partial derivative of $log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ can be expressed as

$$\begin{aligned} \frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} &= -\frac{1}{2}(2\boldsymbol{\Sigma}^{-1} - diag\boldsymbol{\Sigma}^{-1}) + \frac{1}{2}\left[2\boldsymbol{\Xi} - diag\boldsymbol{\Xi}\right] \\ &= \frac{1}{2} diag\left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Xi}\right] - \left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Xi}\right]. \end{aligned} \tag{69}$$

Now denote the matrix $\boldsymbol{S}$ as

$$\boldsymbol{S} \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \int P(\boldsymbol{\theta}|\boldsymbol{y}_{ij})(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)' d\boldsymbol{\theta}. \tag{70}$$

Following the same procedure given by Mislevy (1984, p. 366),

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2} diag\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1}\right] \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{\Sigma}^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij}. \tag{71}$$

Set the above expression to 0, and then $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$. Substituting $\hat{\boldsymbol{\Sigma}}$ into 70 and further simplify the equation in 70 will yield

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \int P(\boldsymbol{\theta}|\boldsymbol{y}_{ij})(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)' d\boldsymbol{\theta} \\ &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \int P(\boldsymbol{\theta}|\boldsymbol{y}_{ij})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}_\theta)' d\boldsymbol{\theta} \\ &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} \tilde{\boldsymbol{\Sigma}}_{ij} + \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{ij} (\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{\mu}_\theta)(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{\mu}_\theta)'. \end{aligned} \tag{72}$$

Therefore, the MML estimates for $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\tilde{\boldsymbol{\Sigma}}_{ij} + \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{\mu}_\theta)(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{\mu}_\theta)'}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}}$$
$$= \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}\tilde{\boldsymbol{\Sigma}}_{ij} + \sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{\Gamma}'\boldsymbol{x}'_{ij} - \boldsymbol{U}'_j\boldsymbol{x}'_{ij})(\tilde{\boldsymbol{\theta}}_{ij} - \boldsymbol{\Gamma}'\boldsymbol{x}'_{ij} - \boldsymbol{U}'_j\boldsymbol{x}'_{ij})'}{\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{ij}}. \quad (73)$$