Research Report

Parameter Recovery and Subpopulation Proficiency Estimation in Hierarchical Latent Regression Models

Deping Li

Andreas Oranje

Yanlin Jiang

Research & Development

June 2007
RR-07-27

Parameter Recovery and Subpopulation Proficiency Estimation in Hierarchical Latent Regression Models

Deping Li, Andreas Oranje, and Yanlin Jiang
ETS, Princeton, NJ

June 2007

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

The hierarchical latent regression model (HLRM) is a flexible framework for estimating group-level proficiency while taking into account the complex sample designs often found in large-scale educational surveys. A complex assessment design in which information is collected at different levels (such as student, school, and district), the model also provides a mechanism for estimating group differences at various levels and for partitioning variance components among those levels. This study examines parameter recovery in the HLRM and compares it to regular latent regression models (LRMs) through simulation for various levels of cluster variation. Results show that regression effect estimates are similar between the HLRM and the LRM, in particular under small cluster variation. Similarly, student posterior mean estimates and marginal maximum likelihood mean estimates for student groups are comparable across the two model approaches. However, substantial differences are found for the residual variance estimates, the standard errors for regression effect estimates and related standard errors for group estimates, and for students posterior variance estimates. As expected, these differences are larger when the variation across clusters is larger, since a substantial portion of variance remains unexplained in LRM.

Key words: Hierarchical latent regressions, item response theory, latent regressions, marginal parameters estimation

Table of Contents

1 Introduction

Latent regression models (LRM) coupled with item response theory (IRT; Mislevy, 1984, 1985) are widely used in large-scale educational survey assessments such as the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS). The frameworks in these low-stakes assessments are broad and represented by a large item pool, while the time available for testing students is limited. As a result, such assessments are typically designed so that test takers each respond to a small portion of the item pool, but, as a group, respond to the total item pool. Hence, reliable estimates of proficiency are limited to the group level after appropriate horizontal equating has been conducted, usually through IRT models. Latent regressions are used to predict the expectations of students' latent abilities from a complicated multistage stratification sample, using a collection of background variables as predictors. In current applications (e.g., Allen, Donoghue, & Schoeps, 2001), test takers are treated as if they were selected from a simple random sample. As a result, is that standard errors for regression effect estimates are likely to be underestimated.

Since estimation of sampling variability of any statistic must take into account the sample design (Cochran, 1977), concerns have been voiced about the use of standard errors based on a simple random sample assumption and how this affects standard errors of group mean estimates (e.g., Cohen & Jiang, 2002; von Davier & Sinharay, 2005). Under NAEP's sampling design, students nested in schools are sampled, constituting a relatively strong cluster effect. This is because, on average, 20 to 25 students are selected from each sampled school, and students within a school share curriculum and instruction experiences. Consequently, by ordinary LRM, it is not possible to adequately capture the population estimates and their variability from a clustered sample. In practice, the population estimates are deemed to be relatively robust against violations of the assumption of independent observations, and post hoc techniques are applied to infer correct standard errors, either via resampling methods (e.g., Johnson & Rust, 1992) or linearization (Cohen & Jiang). However, the viability of these techniques has not been established empirically, likely for lack of an appropriate hierarchical model for comparison.

The hierarchical linear model proposed by Raudenbush and Bryk (2002) has proven to be a useful and effective approach, appropriate for analyzing data when observations within clusters are correlated. Based on this work, Li and Oranje (2006) suggested a two-level latent regression model coupled with IRT to analyze large-scale survey assessment data, where the LRM can be

considered a special case of the more general HLRM. One advantage of this multilevel model is that within-cluster correlations and across-cluster variation can be appropriately accounted for by introducing random effects. Furthermore, under this approach, variability can be specified in terms of variability of individual students within and across schools. Extension of the two-level model to a more general and complicated HLRM is straightforward.

The current study is designed to examine how well parameters from an HLRM can be recovered and how well this approach compares with the current LRM approach using simulation. Additionally, the study examines how well subpopulation characteristics are estimated in the HLRM (i.e., group mean estimates and their standard errors) and how well the estimates of group effect parameters compare with those in the LRM. In the following sections, the family of HLRMs and the estimation of HLRM parameters are briefly introduced. Subsequently, the simulation study design is discussed, and the study results are examined. The final section is devoted to discussion and conclusions.

2 Hierarchical Latent Regression Model

The multilevel IRT model proposed by Li and Oranje (2006) has a fixed-effect model at Level 1 and an unconditional model at Level 2. That is, Level 1 incorporates a model that would otherwise be used in the regular LRM, while Level 2 includes random effects along with fixed effects. In this model, fixed regression effects represent the average of effects across clusters. This formulation was chosen because it allows comparison of an HLRM to current implementations of the regular LRM, where the fixed effects might be based on principal component scores instead of group indicators themselves, as part of a data-reduction effort. Consequently, the resulting regression effect estimates from the two-level HLRM can directly be compared to LRM estimates, while the HLRM approach provides additional information: (a) the variance within clusters and (b) the variance and covariance of random effects across clusters.

In this model, random effects across clusters are assumed and variation across clusters can be characterized by a variance matrix $\boldsymbol{T}$, through which the correlation between two students within each cluster can be calculated. Since the discussion of hierarchical models involves sampling clusters, the notation of clusters will be used to indicate the hierarchical or nested data structure. Suppose there are n_j students nested within J schools for $j = 1, \cdots, J$, and their proficiencies (θ_{1j} to $\theta_{n_j j}$) are likely to be positively correlated. Let $\boldsymbol{x}_{ij}$ be the vector of Q population group

indicators; 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 are denoted as $\boldsymbol{\gamma}_j = (\gamma_{j1}, \cdots, \gamma_{jQ})'$, and the fixed regression effects (or the overall regression effects across clusters) are denoted by $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_Q)'$. The item responses for student i in school j are denoted by $\boldsymbol{y}_{ij}$. Therefore, the Level 1 model is

$$\theta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\gamma}_j + \varepsilon_{ij}, \tag{1}$$

where the residual term ε_{ij} is assumed to be *i.i.d.*, $\varepsilon_{ij} \sim N(0, \sigma^2)$. The Level 2 unconditional model is given by

$$\boldsymbol{\gamma}_j = \boldsymbol{\gamma} + \boldsymbol{u}_j. \tag{2}$$

Substituting (2) into (1) gives the combined model

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

The marginal variance of θ_{ij} is $\boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x}'_{ij} + \sigma^2$, where $\boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x}'_{ij}$ is the variance component associated with the random effects $\boldsymbol{u}_j$ and attributable to the variation across schools. The residual variance σ^2 depicts the variation among students within schools. In this model, variation is decomposed into a school-level and a student-level component. Suppose a student with latent ability $\theta_{i'}$ comes from the same school j as student i, then the covariance of student i and i' can be written as

$$Cov(\theta_{ij}, \theta_{i'j}) \quad = \quad \boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x}_{i'j}, \tag{4}$$

and their correlation can be expressed as

$$Cor(\theta_{ij}, \theta_{i'j}) \quad = \quad \frac{\boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x}_{i'j}}{\sqrt{(\boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x}'_{ij} + \sigma^2)(\boldsymbol{x}_{i'j}\boldsymbol{T}\boldsymbol{x}'_{i'j} + \sigma^2)}}. \tag{5}$$

For example, suppose two students are in the same school and their group indicators are given as $\boldsymbol{x}_{ij} = (1, 0, 0, 0, 1, 1)$ and $\boldsymbol{x}_{i'j} = (0, 1, 1, 0, 0, 0)$, respectively. Suppose also that the diagonal entries of the $\boldsymbol{T}$ matrix are all .05 and off-diagonal entries are all .02, while the residual variance $\sigma^2 = .5$. In that case, the covariance between these two students is .12 and the correlation between them is .17. For the same two students and the same value of σ^2, changing the diagonal components of the $\boldsymbol{T}$ matrix to .5 and the off-diagonal components to .2 (to represent a relatively large variance and covariance across clusters) changes the covariance to 1.20, and the correlation between them changes to .49. Similarly for the same two students, changing the diagonal components of $\boldsymbol{T}$

to .005 and the off-diagonal components to .002 (to represent a relatively small variance and covariance across clusters) changes the covariance to .012, and the correlation between them becomes .023.

The correlations will be larger when two students' background variables share more common values, which is often the case for two students in the same school. Note that when all the components of the $\boldsymbol{T}$ matrix equal zero, the HLRM simplifies to the regular LRM. This implies that there is no variation across clusters or there are no random effects across clusters. In other words, all clusters in the concerned population are homogenous to one another, and the group indicators do not have differential effects across these clusters or schools. In this sense, the HLRM becomes a fixed latent regression model, and the LRM is a special case of the HLRM.

3 Parameter Estimation of HLRM

The procedures for estimating parameters in the HLRM have been developed using marginal maximum likelihood (MML) methods. In the univariate case, the parameters $\boldsymbol{\gamma}, \sigma^2$, and $\boldsymbol{T}$ are estimated from test-response data and student-group indicators. Li and Oranje (2006) developed the MML estimates for these parameters, which involve sampling weights for clusters. A brief summarization of the MML equations for the univariate case will be given in this section (see Li & Oranje, 2006, for details and multivariate extensions); more details on this estimation procedure can be seen in Appendix A.

Suppose the cluster random effects are $\boldsymbol{u}_j$ for $j = 1, \cdots, J$. The *log* likelihood function L over all individual students and test item responses is

$$\begin{aligned} L &= log\left(\prod_{j=1}^{J} P(\boldsymbol{y}_j|\boldsymbol{\theta}_j, \boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j, \sigma^2, \boldsymbol{T})^{w_j}\right) \\ &= \sum_{j=1}^{J} w_j log\left(\int P(\boldsymbol{y}_j|\boldsymbol{\theta}_j, \boldsymbol{u}_j)\phi(\boldsymbol{\theta}, \boldsymbol{u}_j|\boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j, \sigma^2, \boldsymbol{T})d\boldsymbol{\theta}_j d\boldsymbol{u}_j\right) \end{aligned} \quad (6)$$

where ϕ represents the normal density and w_j represents the sampling weights for each school or cluster. An EM algorithm (Dempster, Laird, & Rubin, 1977) can be used for estimation, where in the maximization step the parameters $\boldsymbol{\gamma}$, σ^2, and $\boldsymbol{T}$ given the data $(\boldsymbol{x}_{ij}, \boldsymbol{y}_{ij})$ can be obtained through

$$\hat{\gamma} = \left(\sum_{j=1}^{J} w_j \sum_{i=1}^{n_j} \boldsymbol{x}'_{ij}\boldsymbol{x}_{ij}\right)^{-1} \sum_{j=1}^{J} w_j \sum_{i=1}^{n_j} \boldsymbol{x}'_{ij}(\tilde{\theta}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{u}_j), \quad (7)$$

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{J} w_j \sum_{i=1}^{n_j} \tilde{\sigma}_{ij}^2 + \sum_{j=1}^{J} w_j \sum_{i=1}^{n_j} (\tilde{\theta}_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\gamma}_t - \boldsymbol{x}_{ij}\boldsymbol{u}_j)^2}{N}, \tag{8}$$

$$\hat{\boldsymbol{T}} = \frac{1}{\sum_{j=1}^{J} w_j} \sum_{j=1}^{j} w_j \left[\boldsymbol{C}_{jt}^{-1}\sigma^2 + \tilde{\boldsymbol{u}}_j\tilde{\boldsymbol{u}}_j' + \boldsymbol{C}_{jt}^{-1}\boldsymbol{X}_j' \left[\tilde{\sigma}_{ij}^2\right] \boldsymbol{X}_j\boldsymbol{C}_{jt}^{-1} \right]. \tag{9}$$

where $\tilde{\theta}_{ij}$ in equation (7) represents the posterior mean of a student, and $\tilde{\sigma}_{ij}^2$ in (8) and (9) represents the posterior variance. Hence, during the expectation step of the algorithm, posterior moments need to be evaluated. The posterior expectation of the random effects, $\tilde{\boldsymbol{u}_j}$, is given by

$$\tilde{\boldsymbol{u}_j} = \boldsymbol{C}_j^{-1}\boldsymbol{X}_j'(\tilde{\boldsymbol{\theta}}_j - \boldsymbol{X}_j\boldsymbol{\gamma}), \tag{10}$$

where $\boldsymbol{C}_j$ is expressed by

$$\boldsymbol{C}_j = \boldsymbol{X}_j'\boldsymbol{X}_j + \sigma^2\boldsymbol{T}^{-1}, \tag{11}$$

and $\boldsymbol{X}_j = (\boldsymbol{x}_{ij}', \cdots, \boldsymbol{x}_{n_jj}')'$ and $\tilde{\boldsymbol{\theta}}_j = (\tilde{\theta}_{ij}, \cdots, \tilde{\theta}_{n_jj})'$. The posterior moments for θ_{ij} can be found following

$$\tilde{\theta}_{ij} = \int \theta_{ij} P(\theta_{ij}|\boldsymbol{y}_{ij}) d\theta_{ij} \tag{12}$$

and

$$\tilde{\sigma}_{ij}^2 = \int (\theta_{ij} - \tilde{\theta}_{ij})^2 P(\theta_{ij}|\boldsymbol{y}_{ij}) d\theta_{ij}. \tag{13}$$

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

$$P(\theta_{ij}|\boldsymbol{Y}) = \frac{P(\boldsymbol{y}_{ij}|\theta_{ijt})\phi(\theta_{ij}|\boldsymbol{x}_{ij}\boldsymbol{\gamma}, \boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x'}_{ij} + \sigma^2)}{\int P(\boldsymbol{y}_{ij}|\theta_{ij})\phi(\theta_{ij}|\boldsymbol{x}_{ij}\boldsymbol{\gamma}, \boldsymbol{x}_{ij}\boldsymbol{T}\boldsymbol{x'}_{ij} + \sigma^2) d\theta_{ij}}. \tag{14}$$

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

$$\boldsymbol{T} = \begin{pmatrix} \tau_{11} & \tau_{12} & \cdots & \tau_{1Q} \\ \tau_{21} & \tau_{22} & \cdots & \tau_{2Q} \\ \cdots & \cdots & \cdots & \cdots \\ \tau_{Q1} & \tau_{Q2} & \cdots & \tau_{QQ} \end{pmatrix},$$

where the diagonal element τ_{qq} indicates the variance of random regression effects γ_{qj} for $q = 1, \cdots, Q$ and across all schools $j = 1, \cdots, J$ and the off-diagonal elements $\tau_{qq'}$ indicate the covariance between two random effects (γ_{qj} and $\gamma_{q'j}$) for $q, q' = 1, \cdots, Q$. Appendix A provides more details on how to obtain the estimates for the entries in the matrix $\boldsymbol{T}$.

4 Standard Errors

Standard errors for regression-effect estimates are often approximated by the posterior variance of the regression effects estimates, which can in turn be approximated by the sum of the variance resulting from sampling errors and the variance resulting from the latency of students' abilities. Li and Oranje (2006) followed the derivations by Raudenbush and Bryk (2002, pp. 44–45) and suggested the generalized least squares estimator for fixed effects in the HLRM as

$$\hat{\gamma} = \left(\sum_{j=1}^{J} \boldsymbol{X}_j' \boldsymbol{D}_j \boldsymbol{V}_{\boldsymbol{\theta}_j}^{-1} \boldsymbol{X}_j \right)^{-1} \sum_{j=1}^{J} \boldsymbol{X}_j' \boldsymbol{D}_j \boldsymbol{V}_{\boldsymbol{\theta}_j}^{-1} \boldsymbol{\theta}_j, \tag{15}$$

where

$$V_{\boldsymbol{\theta}_j} = Var(\boldsymbol{\theta}_j) = \boldsymbol{X}_j \boldsymbol{T} \boldsymbol{X}_j' + \sigma^2 \boldsymbol{I}, \tag{16}$$

and D_j is a diagonal matrix of student sampling weights. Hence, the variance covariance matrix for fixed effect estimates $\hat{\gamma}$ is

$$Var(\hat{\gamma}) = \left(\sum_{j=1}^{J} \boldsymbol{X}_j' \boldsymbol{D}_j \boldsymbol{V}_{\boldsymbol{\theta}_j}^{-1} \boldsymbol{X}_j \right)^{-1}. \tag{17}$$

Using (17) to compute standard errors for the fixed effects assumes that student abilities are observed. In the HLRM, the variation due to the latency of the construct should be added; hence, the variance $Var(\hat{\gamma})$ is given by

$$Var(\hat{\gamma}) = \boldsymbol{V} \left[\sum_{j=1}^{J} \boldsymbol{X}_j' \boldsymbol{D}_j \boldsymbol{V}_{\theta_j}^{-1} (\boldsymbol{V}_{\theta_j} + \tilde{\boldsymbol{\Sigma}}_j) \boldsymbol{V}_{\theta_j}^{-1} \boldsymbol{D}_j \boldsymbol{X}_j \right] \boldsymbol{V}, \tag{18}$$

where $\boldsymbol{V} = \left(\sum_{j=1}^{J} \boldsymbol{X}_j' \boldsymbol{D}_j \boldsymbol{V}_{\theta_j}^{-1} \boldsymbol{X}_j \right)^{-1}$. More details on the derivation of the standard error estimates in the HLRM can be found in Section 3 in Li and Oranje (2006).

5 Subpopulation Characteristics Estimates

The HLRM is an efficient approach for obtaining accurate group-ability estimates from test-response data, which is of primary interests for reporting of test results. The MML estimate of the average proficiency for group G can be derived from the regression coefficient estimates (Mazzeo, Donoghue, Li, & Johnson, 2005):

$$\hat{\boldsymbol{\mu}}_G = \hat{\boldsymbol{\Gamma}}' \bar{\boldsymbol{x}}_G', \tag{19}$$

where $\bar{\boldsymbol{x}}'_G$ is the sample mean vector of the background variables across examinees in group G,

$$\bar{\boldsymbol{x}}'_G = \frac{\sum_{i \in G} w_i \boldsymbol{x}_i}{\sum_{i \in G} w_i}. \tag{20}$$

The variance of the group mean estimate can be computed by

$$Var(\hat{\boldsymbol{\mu}}_G) = \bar{\boldsymbol{x}}_G \, Var(\hat{\boldsymbol{\Gamma}}) \bar{\boldsymbol{x}}'_G. \tag{21}$$

The empirical Bayes estimate of the same quantity is given by

$$\tilde{\boldsymbol{\mu}}_G = \sum_{i \in G} \frac{w_i \tilde{\mu}_i}{\sum_{i \in G} w_i}, \tag{22}$$

where $\tilde{\mu}_i$ is the mean for the posterior distribution for examinee i evaluated at $\hat{\boldsymbol{\Gamma}}$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{T}}$.

6 Simulation Design

In this simulation study, a 50-item test is generated under the 3PL item-response model (e.g., Lord, 1980) with discrimination parameters distributed as $a_i \sim U(.5, 1.5)$, difficulty parameters as $b_i \sim N(0, 1)$, and asymptote parameters as $c_i \sim U(0, .4)$ for $i = 1, \cdots, n$. The realizations can be found in Table 1. Two conditions for simulees and schools are employed with 2,000 and 5,000 simulees within 100 and 250 schools, respectively. That is, each school has 20 students whose latent abilities are conditionally normal distributed and positively correlated. The correlation of latent abilities of students from the same schools can be calculated from (5). These sample sizes characterize the range of sample sizes for a typical state assessment in the National Assessment of Educational Progress (NAEP).

In this study, the number of items n, the regression parameters $\boldsymbol{\gamma}$ and σ^2, and the item parameters are all fixed, with $\boldsymbol{\gamma} = (.25, -.12, -.83, -.63, .23, .41)'$ for six student group indicators (predictors) and $\sigma^2 = 0.5$. Hence, only two variables are designed to vary: (a) the variance of random effects $\boldsymbol{T}$ and (b) the sample size. For a given sample size, the only difference between the HLRM and the LRM is the size of the cluster effects, which are assumed to be random effects in the HLRM and fixed in the LRM with the result that $\boldsymbol{T}$ becomes a matrix with zeros.

As was noted in previous sections, the covariance matrix $\boldsymbol{T}$ of random effects across clusters reflects the magnitude of the correlation between examinees within the same schools. The bigger the magnitude of the diagonal entries in the matrix $\boldsymbol{T}$, the larger the variation across clusters and the larger the correlations between students within the same schools. Furthermore, the larger the

Table 1

True Item Parameters

Item	Discrimination a	Difficulty b	Asymptote c
1	0.953	-1.489	0.06
2	1.305	-0.128	0.131
3	0.684	-0.546	0.209
4	0.791	-0.898	0.213
5	0.696	1.535	0.067
6	1.322	-0.4	0.308
7	1.29	1.122	0.187
8	1.246	-0.713	0.262
9	0.736	0.649	0.269
10	0.802	1.016	0.11
11	0.989	-0.977	0.164
12	1.127	-1.162	0.178
13	1.21	-0.995	0.023
14	0.629	0.538	0.305
15	1.037	-0.449	0.339
16	0.66	0.722	0.181
17	0.615	-1.981	0.249
18	0.683	-1.018	0.072
19	0.802	0.132	0.298
20	1.399	0.381	0.065
21	1.349	0.362	0.259
22	0.609	-0.277	0.167
23	0.945	-1.295	0.232
24	1.315	0.731	0.241
25	1.346	1.208	0.25
26	1.256	-0.082	0.348
27	1.086	0.45	0.307
28	1.009	0.738	0.333
29	1.079	0.603	0.035
30	0.697	-0.964	0.311
31	0.897	0.578	0.259
32	1.279	0.535	0.307
33	1.213	-1.467	0.11
34	0.997	0.765	0.254
35	0.98	-0.266	0.011
36	1.325	-2.379	0.107
37	0.976	-0.018	0.143
38	0.847	-1.133	0.295
39	0.943	-1.064	0.341
40	0.683	-0.143	0.243
41	1.216	-0.065	0.266
42	1.146	1.462	0.05
43	0.6	-0.345	0.099
44	0.826	1.092	0.329
45	1.068	-0.028	0.065
46	1.386	0.548	0.335
47	0.618	0.256	0.205
48	0.872	0.883	0.294
49	1.113	-0.39	0.044
50	1.254	1.604	0.154

off-diagonal components are in the matrix $\boldsymbol{T}$, the stronger the associations are between regression effects among clusters. In the simulation, three matrices are used, corresponding to strong, moderate, and weak variation across clusters. Table 2 shows the values used.

Table 2
Design of the T Matrix

	Condition	Diagonal	Off-diagonal
1	Strong	.5	.2
2	Moderate	.05	.02
3	Weak	.005	.002

With six predictors in the latent regression, each $\boldsymbol{T}$ has dimension 6×6. For each condition, 50 replications were conducted, resulting in 150 data sets for each of two sample sizes. The predictors are independent Bernoulli variables where the first predictor is an intercept. These variables are generated to represent typical group distributions such as gender and school lunch eligibility.

For each parameter, the average estimate, the model-based and empirical standard error, and the root mean squared deviation (RMSD) from the true parameter value will be computed to indicate the recovery success. These statistics will be compared with their counterparts based on a regular LRM.

7 Results

7.1 Parameter Recovery

Tables 3 and 4 show the true parameter values, the average parameter estimates, and the average model-based standard error between parentheses using the HLRM and the LRM for all three cluster variation conditions and both sample sizes across 50 replicates. The last three columns show the RMSDs. Regression parameters are generally well recovered across conditions and sample sizes for both models, which is reflected in both the averages and the RMSDs. However, the residual variances are not well recovered under the LRM when variation across clusters is strong or moderate and standard errors seem inadequate. Furthermore, in the strong condition (i.e., Condition 1), RMSDs for regression parameters are substantially larger across both models relative to the moderate and weak conditions, indicating less than desirable estimation

results in general for that condition. This is not surprising as the effective sample size is drastically reduced under the existence of considerable within-cluster dependencies. As expected, estimation size is more accurate with a sample size of 5,000 than a sample size of 2,000, which is reflected in the RMSDs.

Table 3
Parameter Recovery and Standard Errors Over 50 Replications

Size	$\boldsymbol{\gamma}, \sigma^2$	*True*	Mean (average standard error)			RMSD		
			1	2	3	1	2	3
2,000	γ_1	.25	.238 (.076)	.241 (.037)	.251 (.032)	.097	.043	.036
	γ_2	-.12	-.094 (.072)	-.124 (.040)	-.124 (.035)	.126	.051	.040
	γ_3	-.85	-.796 (.080)	-.840 (.048)	-.851 (.044)	.142	.062	.051
	γ_4	-.63	-.576 (.092)	-.620 (.064)	-.633 (.058)	.177	.077	.056
	γ_5	.23	.199 (.073)	.226 (.043)	.232 (.038)	.098	.052	.037
	γ_6	.41	.362 (.072)	.420 (.040)	.406 (.035)	.111	.046	.042
	σ^2	.50	.462 (.020)	.464 (.020)	.472 (.020)	.043	.041	.034
5,000	γ_1	.25	.262 (.048)	.248 (.023)	.256 (.020)	.063	.027	.023
	γ_2	-.12	-.126 (.045)	-.120 (.025)	-.125 (.022)	.055	.028	.021
	γ_3	-.85	-.812 (.050)	-.847 (.031)	-.853 (.027)	.074	.034	.030
	γ_4	-.63	-.580 (.059)	-.614 (.040)	-.632 (.036)	.122	.043	.041
	γ_5	.23	.224 (.047)	.227 (.027)	.235 (.024)	.069	.025	.021
	γ_6	.41	.375 (.046)	.405 (.025)	.406 (.022)	.060	.035	.018
	σ^2	.50	.458 (.010)	.465 (.013)	.487 (.011)	.043	.037	.016

Comparing the RMSDs with the model-based standard errors shows that the model-based standard errors seem on average to be underestimated. This is most noticeable for the residual variance. However, this underestimation is is a result of bias and not the standard errors themselves. To substantiate this finding further, Tables 5 and 6 present the standard deviation (i.e., empirical standard error) for the parameter estimates over 50 replications for each condition, which is in squared terms equivalent to the RMSD minus bias. As the tables show, the standard deviations almost exactly match the model-based standard errors for both models for the residual variances. However, the regression-effect standard errors are somewhat underestimated. Comparing across sample sizes, it appears that the estimates based on 5,000 simulees are less variable than the estimates based on 2,000 simulees, but not necessarily less biased.

Table 7 shows the average estimates for the elements of cluster variation matrix $\boldsymbol{T}$ under the three conditions. In general, the $\boldsymbol{T}$ matrixs seem reasonably well recovered over the three

Table 4
LRM Parameter Recovery and Standard Errors Over 50 Replications

Size	$\boldsymbol{\gamma}, \sigma^2$	True	Mean (Average standard error)			RMSD		
			1	2	3	1	2	3
	γ_1	.25	.240 (.064)	.240 (.037)	.251 (.032)	.095	.043	.036
	γ_2	-.12	-.099 (.069)	-.125 (.040)	-.124 (.034)	.113	.051	.041
	γ_3	-.85	-.792 (.087)	-.841 (.049)	-.852 (.043)	.129	.062	.051
	γ_4	-.63	-.578 (.118)	-.617 (.066)	-.634 (.058)	.156	.076	.056
	γ_5	.23	.200 (.076)	.228 (.043)	.233 (.038)	.090	.051	.037
	γ_6	.41	.365 (.071)	.423 (.040)	.407 (.035)	.104	.046	.042
2,000	σ^2	.50	2.279 (.325)	.702 (.032)	.523 (.018)	1.808	.205	.029
	γ_1	.25	.262 (.041)	.248 (.023)	.256 (.020)	.061	.028	.023
	γ_2	-.12	-.126 (.044)	-.121 (.025)	-.126 (.022)	.052	.028	.021
	γ_3	-.85	-.808 (.055)	-.847 (.032)	-.854 (.027)	.068	.034	.030
	γ_4	-.63	-.578 (.074)	-.614 (.042)	-.632 (.036)	.108	.043	.042
	γ_5	.23	.222 (.048)	.227 (.027)	.235 (.024)	.061	.025	.021
	γ_6	.41	.378 (.045)	.408 (.026)	.407 (.022)	.056	.024	.018
5,000	σ^2	.50	2.306 (.114)	.709 (.020)	.522 (.011)	1.810	.210	.025

conditions. However, it appears that when the cluster variation is strong (i.e., Condition 1), the elements of $\boldsymbol{T}$ tend to be underestimated. When the cluster variation is weak (i.e., Condition 3) the diagonal elements of $\boldsymbol{T}$ tend to be overestimated for the smaller sample size condition. Logically, with more schools the between school variability can be estimated with greater accuracy. More precisely, the accuracy of $\boldsymbol{T}$ depends on the number of clusters J. However, it seems most notably to improve estimation of the diagonal elements. From Table 5, it also becomes clear that for the elements from $\boldsymbol{T}$, the variability associated with parameter estimates is larger in more clustered samples, which is related to the effective sample size.

Special attention should be devoted to the residual variances, especially with respect to the LRM estimates, which are, on average, 2.279, .702, and .523, where the true value is .500. It is obvious that the residual variance estimates from LRM are inadequate and grossly overestimated when the clustering is either moderate or strong. This result is not surprising since the LRM does not account for the cluster variation. The implication, however, is that the variation across clusters is left to the residual variance and, therefore, the estimates for the residual variances are always overestimated in the presence of clustering. The result is that the standard errors under the LRM are overestimated as well, which means that the difference between HLRM and LRM

Table 5
HLRM Empirical Standard Errors Over 50 Replications

	2,000			5,000		
$\boldsymbol{\gamma}, \sigma^2$	1	2	3	1	2	3
γ_1	.098	.042	.037	.062	.027	.022
γ_2	.125	.051	.041	.055	.029	.021
γ_3	.134	.062	.051	.064	.034	.030
γ_4	.170	.077	.057	.112	.041	.042
γ_5	.094	.052	.037	.069	.025	.020
γ_6	.102	.045	.043	.050	.035	.017
σ^2	.020	.020	.020	.010	.013	.011
τ_{11}	.117	.015	.006	.060	.011	.001
τ_{22}	.080	.017	.007	.036	.009	.001
τ_{33}	.088	.025	.010	.050	.015	.001
τ_{44}	.124	.040	.013	.054	.019	.001
τ_{55}	.091	.022	.008	.044	.011	.001
τ_{66}	.083	.019	.006	.047	.012	.001
τ_{12}	.059	.011	.005	.031	.007	.001
τ_{13}	.057	.012	.006	.036	.007	.001
τ_{14}	.074	.017	.006	.042	.006	.001
τ_{15}	.058	.011	.004	.031	.006	.001
τ_{16}	.050	.013	.004	.030	.008	.001
τ_{23}	.064	.013	.005	.030	.008	.001
τ_{24}	.065	.018	.005	.033	.008	.001
τ_{25}	.051	.013	.004	.027	.007	.001
τ_{26}	.050	.013	.004	.029	.009	.001
τ_{34}	.072	.020	.006	.042	.009	.000
τ_{35}	.046	.016	.006	.035	.009	.001
τ_{36}	.050	.014	.006	.030	.009	.001
τ_{45}	.070	.016	.006	.039	.009	.001
τ_{46}	.076	.018	.006	.033	.009	.001
τ_{56}	.055	.014	.004	.030	.007	.001

standard errors is not necessarily a straightforward function of the degree of clustering.

7.2 Subgroup Means and Standard Errors

To investigate typical student group means and standard errors, two groups were constructed using the second variable, $\boldsymbol{x}_2$, from the simulated predictors. Specifically, Group 1 contains simulees with $\boldsymbol{x}_2 = 1$, and Group 2 contains those with $\boldsymbol{x}_2 = 0$. The distribution between these groups is approximately uniform, reflecting a typical gender variable. Tables 8 through **??** present

Table 6
LRM Empirical Standard Errors Over 50 Replications

	2,000			5,000		
$\boldsymbol{\gamma}, \sigma^2$	1	2	3	1	2	3
γ_1	.060	.046	.037	.057	.028	.022
γ_2	.052	.051	.041	.065	.028	.021
γ_3	.055	.062	.051	.064	.034	.030
γ_4	.095	.076	.056	.077	.041	.042
γ_5	.061	.051	.037	.056	.025	.020
γ_6	.046	.044	.043	.069	.034	.017
σ^2	.325	.032	.018	.113	.021	.015

Table 7
HLRM Parameters Recovery Over 50 Replications

	2,000			5,000		
$\boldsymbol{\gamma}, \sigma^2$	1	2	3	1	2	3
True	.500	.050	.005	.500	.050	.005
τ_{11}	.468	.039	.009	.472	.038	.004
τ_{22}	.386	.045	.009	.376	.041	.004
τ_{33}	.394	.046	.014	.382	.044	.004
τ_{44}	.355	.062	.018	.389	.049	.004
τ_{55}	.364	.042	.011	.369	.043	.004
τ_{66}	.374	.039	.009	.383	.041	.004
True	.200	.020	.002	.200	.020	.002
τ_{12}	.156	.012	-.003	.155	.013	.001
τ_{13}	.120	.009	-.003	.126	.014	.001
τ_{14}	.117	.010	-.001	.118	.013	.001
τ_{15}	.135	.012	-.002	.138	.014	.001
τ_{16}	.140	.012	-.001	.144	.014	.001
τ_{23}	.116	.014	.001	.119	.016	.001
τ_{24}	.108	.014	.001	.111	.016	.001
τ_{25}	.115	.014	.002	.122	.018	.001
τ_{26}	.118	.018	.001	.118	.016	.001
τ_{34}	.111	.016	.001	.116	.016	.001
τ_{35}	.101	.015	.003	.113	.017	.001
τ_{36}	.111	.020	.002	.112	.015	.001
τ_{45}	.100	.013	.001	.115	.019	.001
τ_{46}	.086	.015	.001	.106	.016	.001
τ_{56}	.117	.014	.002	.122	.018	.001

the MML estimates of means and standard errors following (19) and (21), respectively. The statistics for the two groups are based on both the LRM and the HLRM for sample sizes equal to 5,000 and for all three conditions of cluster variation. Similar results were obtained with sample sizes equal to 2,000 and, hence, are not reported in this paper. The results in the three tables show that most of the difference in subgroup mean estimates among the replicated 50 data sets are trivial between the HLRM and the LRM.

Table 8

Subpopulation (x_2) Means Estimates Over Condition 1 (5,000)

	HLRM				LRM			
Data	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE
1	.256	.091	.343	.068	.26	.032	.346	.032
2	.263	.091	.317	.064	.263	.031	.319	.033
3	.17	.091	.353	.065	.17	.032	.353	.032
4	.312	.087	.432	.063	.313	.031	.436	.03
5	.144	.085	.237	.065	.144	.031	.238	.031
6	.266	.09	.352	.064	.265	.032	.354	.031
7	.124	.089	.283	.065	.124	.031	.285	.031
8	-.024	.087	.203	.065	-.018	.032	.195	.031
9	.194	.088	.276	.064	.19	.031	.283	.031
10	.081	.095	.253	.066	.088	.033	.245	.032
11	.154	.089	.312	.067	.156	.032	.313	.031
12	-.029	.089	.158	.066	-.026	.032	.156	.032
13	.033	.09	.279	.064	.041	.031	.273	.031
14	.109	.091	.246	.066	.11	.032	.248	.032
15	.123	.09	.211	.067	.121	.032	.216	.032
16	.105	.087	.246	.06	.106	.03	.245	.031
17	.016	.089	.137	.066	.013	.032	.141	.032
18	.152	.087	.202	.061	.149	.031	.208	.03
19	.254	.086	.338	.063	.256	.031	.339	.031
20	.073	.095	.221	.068	.077	.033	.22	.033
21	.051	.085	.26	.06	.051	.031	.263	.03
22	.198	.085	.303	.06	.2	.03	.3	.03
23	.233	.087	.355	.064	.234	.032	.354	.031
24	.159	.087	.319	.064	.161	.031	.322	.031
25	.142	.084	.314	.059	.144	.03	.316	.03
26	.014	.089	.203	.067	.012	.032	.208	.031
27	.18	.088	.243	.062	.18	.031	.244	.031
28	.352	.084	.393	.06	.351	.03	.398	.03

Table continues

Table 8 (continued)

	HLRM				LRM			
Data	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE
29	.194	.088	.253	.062	.191	.031	.26	.031
30	.085	.084	.253	.061	.091	.031	.249	.03
31	.1	.086	.258	.06	.099	.031	.264	.03
32	.162	.089	.283	.061	.159	.031	.291	.031
33	.035	.088	.312	.065	.039	.031	.307	.031
34	.057	.087	.266	.062	.062	.031	.263	.031
35	.066	.096	.204	.068	.072	.033	.202	.033
36	.001	.088	.216	.065	.004	.031	.216	.031
37	.12	.084	.257	.06	.124	.03	.254	.03
38	.146	.09	.288	.065	.149	.032	.288	.031
39	.187	.089	.335	.062	.191	.031	.331	.031
40	.186	.092	.285	.065	.188	.032	.287	.032
41	.185	.093	.269	.067	.183	.032	.274	.032
42	.039	.087	.245	.066	.041	.031	.242	.031
43	.152	.09	.293	.063	.153	.032	.296	.031
44	.104	.085	.241	.06	.103	.03	.246	.031
45	-.059	.084	.179	.062	-.06	.03	.18	.03
46	.186	.088	.292	.065	.183	.031	.299	.032
47	.186	.083	.328	.057	.184	.03	.337	.03
48	.125	.093	.224	.068	.124	.032	.226	.033
49	.013	.086	.211	.063	.013	.031	.212	.031
50	.138	.092	.221	.067	.137	.033	.227	.031

Table 9

Subpopulation (x_2) Means Estimates Over Condition 2 (5,000)

	HLRM				LRM			
Data	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE
1	.103	.029	.279	.024	.103	.017	.28	.018
2	.128	.035	.246	.026	.128	.018	.246	.019
3	.189	.031	.309	.026	.19	.018	.311	.018
4	.104	.034	.185	.024	.105	.018	.186	.018
5	.14	.031	.269	.023	.14	.018	.27	.017
6	.188	.034	.257	.025	.186	.018	.258	.018
7	.113	.031	.256	.024	.114	.017	.256	.018
8	.137	.033	.3	.025	.137	.018	.3	.018

(Table continues)

Table 9 (continued)

	HLRM				LRM			
Data	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE
9	.115	.033	.212	.024	.116	.018	.212	.018
10	.144	.031	.259	.024	.147	.018	.258	.018
11	.105	.032	.28	.024	.105	.018	.281	.018
12	.195	.033	.314	.025	.196	.018	.316	.018
13	.145	.032	.249	.023	.145	.018	.25	.018
14	.161	.034	.251	.026	.161	.018	.252	.018
15	.05	.034	.191	.026	.049	.019	.192	.018
16	.104	.031	.261	.022	.104	.018	.263	.018
17	.105	.032	.25	.024	.105	.018	.252	.018
18	.105	.031	.252	.024	.106	.018	.255	.018
19	.119	.032	.29	.024	.12	.018	.29	.018
20	.194	.032	.305	.025	.195	.018	.307	.018
21	.11	.031	.231	.026	.108	.018	.233	.018
22	.131	.031	.291	.023	.132	.018	.291	.017
23	.138	.032	.269	.024	.139	.018	.27	.018
24	.164	.031	.281	.025	.164	.018	.281	.018
25	.084	.031	.252	.026	.084	.018	.254	.018
26	.091	.03	.259	.024	.09	.018	.259	.018
27	.152	.034	.293	.026	.153	.018	.294	.018
28	.143	.032	.289	.024	.144	.018	.288	.018
29	.067	.03	.248	.023	.066	.017	.248	.017
30	.1	.032	.252	.022	.101	.018	.252	.018
31	.131	.031	.25	.024	.133	.018	.25	.017
32	.079	.032	.283	.024	.082	.018	.284	.018
33	.129	.028	.239	.024	.127	.017	.24	.018
34	.151	.031	.249	.023	.151	.018	.249	.017
35	.042	.03	.222	.022	.04	.017	.222	.017
36	.162	.031	.282	.024	.161	.017	.284	.018
37	.112	.029	.24	.024	.111	.018	.24	.017
38	.139	.033	.293	.026	.139	.018	.295	.018
39	.123	.03	.248	.023	.123	.017	.248	.018
40	.169	.032	.27	.023	.168	.018	.271	.018
41	.133	.031	.244	.026	.133	.018	.245	.018
42	.042	.031	.219	.024	.043	.018	.218	.017
43	.106	.031	.235	.025	.107	.018	.236	.018
44	.114	.032	.249	.026	.115	.018	.248	.018
45	.14	.033	.282	.024	.14	.018	.282	.018
46	.177	.03	.305	.025	.176	.018	.307	.018
47	.131	.032	.254	.024	.131	.018	.254	.018
48	.122	.032	.231	.025	.121	.018	.231	.018
49	.122	.032	.24	.024	.121	.018	.24	.018
50	.116	.032	.267	.024	.117	.018	.268	.018

Table 10

Subpopulation (x_2) Means Estimates Over Condition 3 (5,000)

	HLRM				LRM			
Data	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE
1	.144	.017	.276	.016	.143	.015	.277	.015
2	.15	.017	.271	.016	.149	.015	.271	.016
3	.121	.017	.284	.017	.121	.015	.284	.016
4	.142	.018	.282	.016	.142	.015	.282	.015
5	.134	.017	.254	.016	.134	.015	.254	.015
6	.148	.017	.262	.016	.149	.016	.262	.015
7	.144	.017	.283	.016	.144	.016	.283	.016
8	.116	.017	.237	.016	.115	.015	.237	.015
9	.156	.017	.264	.016	.156	.016	.264	.016
10	.123	.017	.269	.016	.123	.016	.269	.015
11	.12	.017	.266	.016	.12	.015	.265	.015
12	.105	.017	.255	.016	.104	.015	.255	.016
13	.109	.016	.286	.015	.108	.015	.286	.015
14	.121	.017	.284	.016	.121	.015	.284	.015
15	.126	.016	.267	.016	.126	.015	.267	.015
16	.117	.016	.23	.016	.116	.015	.23	.016
17	.086	.018	.228	.016	.086	.016	.228	.016
18	.132	.017	.236	.016	.133	.016	.236	.015
19	.169	.017	.284	.016	.168	.016	.285	.015
20	.109	.017	.269	.016	.109	.016	.27	.016
21	.123	.017	.302	.016	.123	.016	.301	.015
22	.144	.017	.248	.016	.143	.015	.248	.015
23	.138	.017	.268	.016	.137	.016	.268	.015
24	.126	.017	.298	.016	.125	.015	.298	.015
25	.125	.017	.247	.016	.125	.015	.247	.015
26	.093	.018	.252	.016	.094	.016	.25	.015
27	.109	.018	.283	.016	.108	.016	.284	.016
28	.144	.016	.261	.016	.144	.015	.26	.015
29	.138	.018	.286	.016	.137	.016	.286	.015
30	.145	.017	.262	.015	.144	.015	.262	.015
31	.124	.016	.261	.015	.123	.015	.262	.015
32	.119	.018	.261	.017	.118	.016	.261	.016
33	.156	.017	.288	.016	.155	.015	.289	.015
34	.135	.017	.276	.016	.134	.015	.277	.015
35	.126	.017	.267	.017	.126	.015	.267	.016
36	.113	.017	.26	.016	.114	.015	.26	.015

(Table continues)

Table 10 (continued)

	HLRM				LRM			
Data	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE	$\boldsymbol{x}_2 = 1$	SE	$\boldsymbol{x}_2 = 0$	SE
37	.109	.017	.263	.016	.108	.015	.264	.015
38	.141	.017	.248	.016	.14	.016	.248	.015
39	.115	.017	.278	.016	.114	.015	.278	.015
40	.127	.017	.267	.016	.127	.015	.267	.015
41	.127	.016	.295	.016	.127	.015	.294	.015
42	.115	.016	.265	.016	.114	.015	.266	.015
43	.125	.017	.283	.016	.125	.016	.283	.016
44	.138	.017	.266	.016	.139	.015	.265	.016
45	.108	.018	.252	.016	.109	.015	.252	.015
46	.139	.017	.282	.016	.139	.015	.282	.016
47	.133	.016	.283	.016	.132	.015	.283	.015
48	.137	.018	.245	.016	.137	.016	.245	.016
49	.104	.018	.272	.017	.103	.016	.272	.016
50	.118	.017	.255	.016	.118	.016	.255	.015

The standard errors are different between the two models. Where the LRM only shows moderate increase in the standard errors with increasing clustering, mainly due to inflation as a result of imprecise residual variance estimation, the HLRM increase is substantial, seemingly appropriately accounting for the complex sample design. The reason is that, in (18), $\boldsymbol{T}$ contributes to the standard error estimation in the HLRM, outweighing the inflation in the LRM approach. As expected, under the weak intracluster correlation condition, estimates between both models are largely equivalent.

The MML subgroup mean estimates are directly affected by the regression effect estimates, which are in turn functions of the posterior mean estimates. The fact that the group means are similar between the HLRM and the LRM implies that posterior mean estimates are similar across the two models. Figures 1 through 3 are plots of the regression parameter estimates between the HLRM and the LRM for 50 replicates, showing that the difference between the models decreases as the cluster variation decreases. Similarly, Figures 4 through 6 are plots of student's posterior mean estimates between the HLRM and the LRM for the first four data sets. As expected, for weak cluster variation the estimates are more similar between the two models than for strong cluster variation.

The standard errors for subgroup means estimates are functions of the variance of the regression effect estimates, which are in turn related to the residual variance estimate and the

posterior variance estimates for all students. Figures 7 through 9 show the plots of each students' posterior variance estimate between the HLRM and the LRM across the three different conditions of cluster variation for the first four data sets. It appears that, similar to the posterior means, for strong cluster variation students' posterior variance estimates are quite different between the two models but similar when this variation is weak. When they are different, they are larger for the LRM because the posterior means are estimated less accurately.

8 Conclusion and Discussion

In this simulation study, an HLRM has been evaluated for parameter recovery and compared against a regular LRM. The study provides some empirical evidence to support several conclusions about HLRM parameter estimation. Compared to the LRM, the HLRM explicitly provides a mechanism to account for the variation across clusters (i.e., the variance matrix $\boldsymbol{T}$), which is often substantial in educational survey assessments. Models such as the LRM often ignore the clustering, probably due to lack of an appropriate model, resulting in overestimation of residual variances.

In general, regression effect estimates are very similar between the HLRM and the LRM for a wide range of clustering levels such as those employed in this study. for other parameters, however, some differences emerge when the variation across clusters becomes strong. Similar to hierarchical linear models, the estimates for fixed regression effects are unbiased for any sample size (Raudenbush & Bryk, 2002, pp. 281). The residual variance estimates for the HLRM are reasonably well captured, as reflected in relatively small root mean squared deviations. However, the residual variances for the LRM are severely overestimated, in particular when the cluster variation is strong. This can be attributed to the fact that the LRM does not account for the cluster variation; hence, the variation across clusters becomes largely part of the residual variance. As a result, the advantage of the HLRM over the LRM with respect to the estimation and interpretation of the variance structure pertains predominantly to the estimation of standard errors.

The components of the cluster variation matrix $\boldsymbol{T}$ seem reasonably well-estimated. They tend to be underestimated when the cluster variation is strong and overestimated when the cluster variation becomes small. However, the degree of over- and underestimation is relatively small. It also should be noted that cluster variation is better estimated with a larger number of clusters.

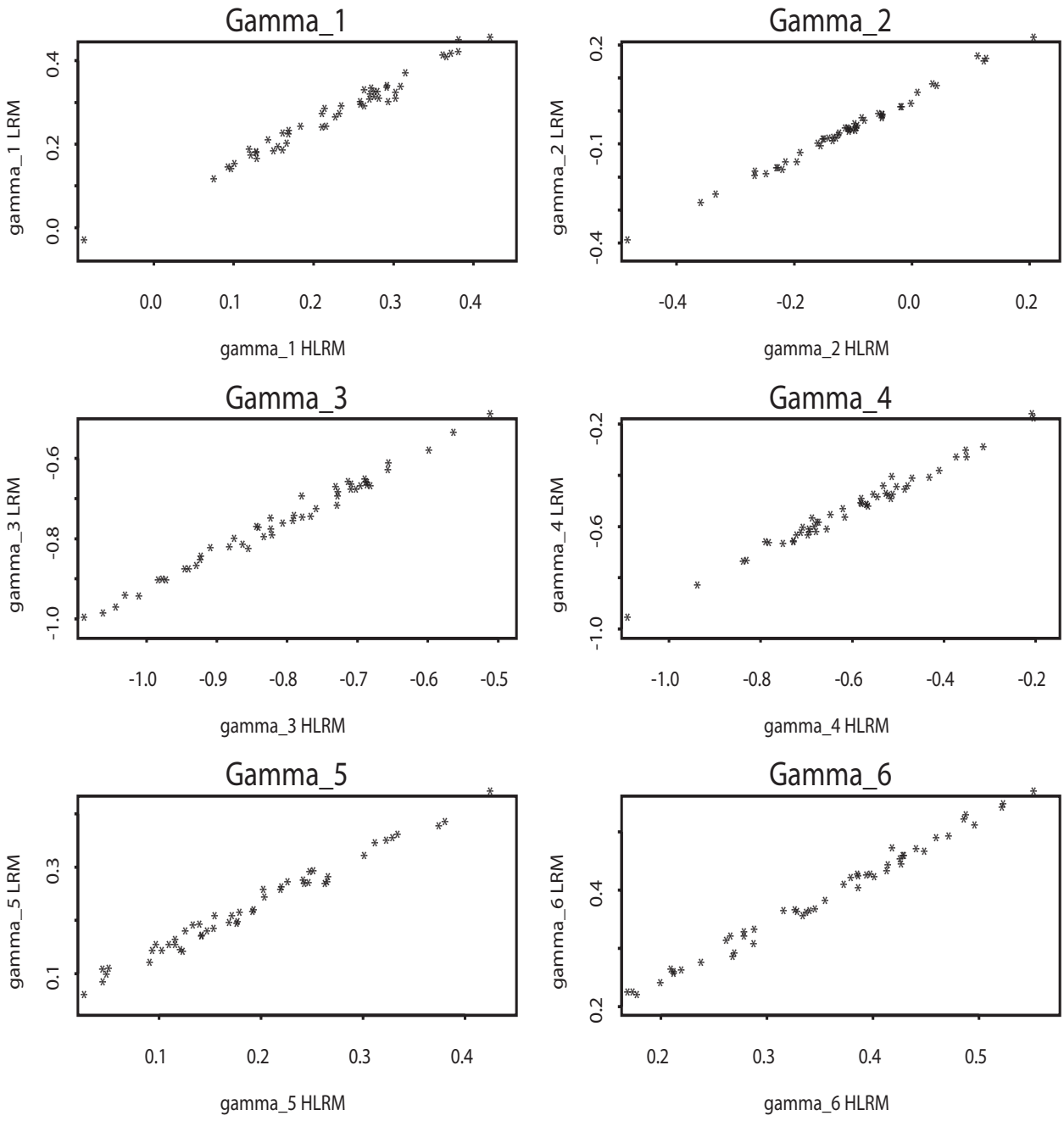

***Figure* 1 Plots of estimates of regression effect parameters over 50 replications in Condition 1.**

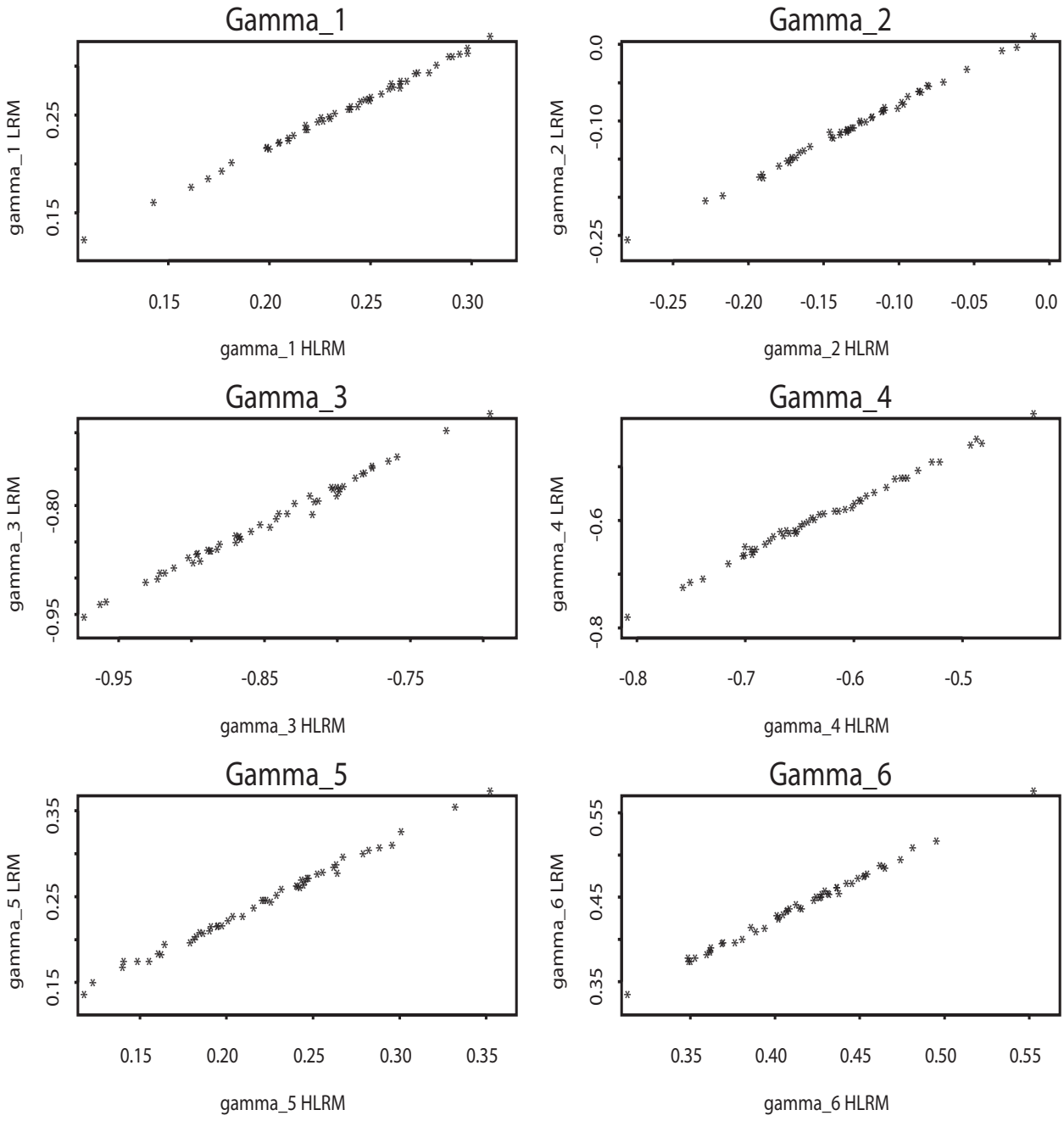

***Figure* 2 Plots of estimates of regression effect parameters over 50 replications in Condition 2.**

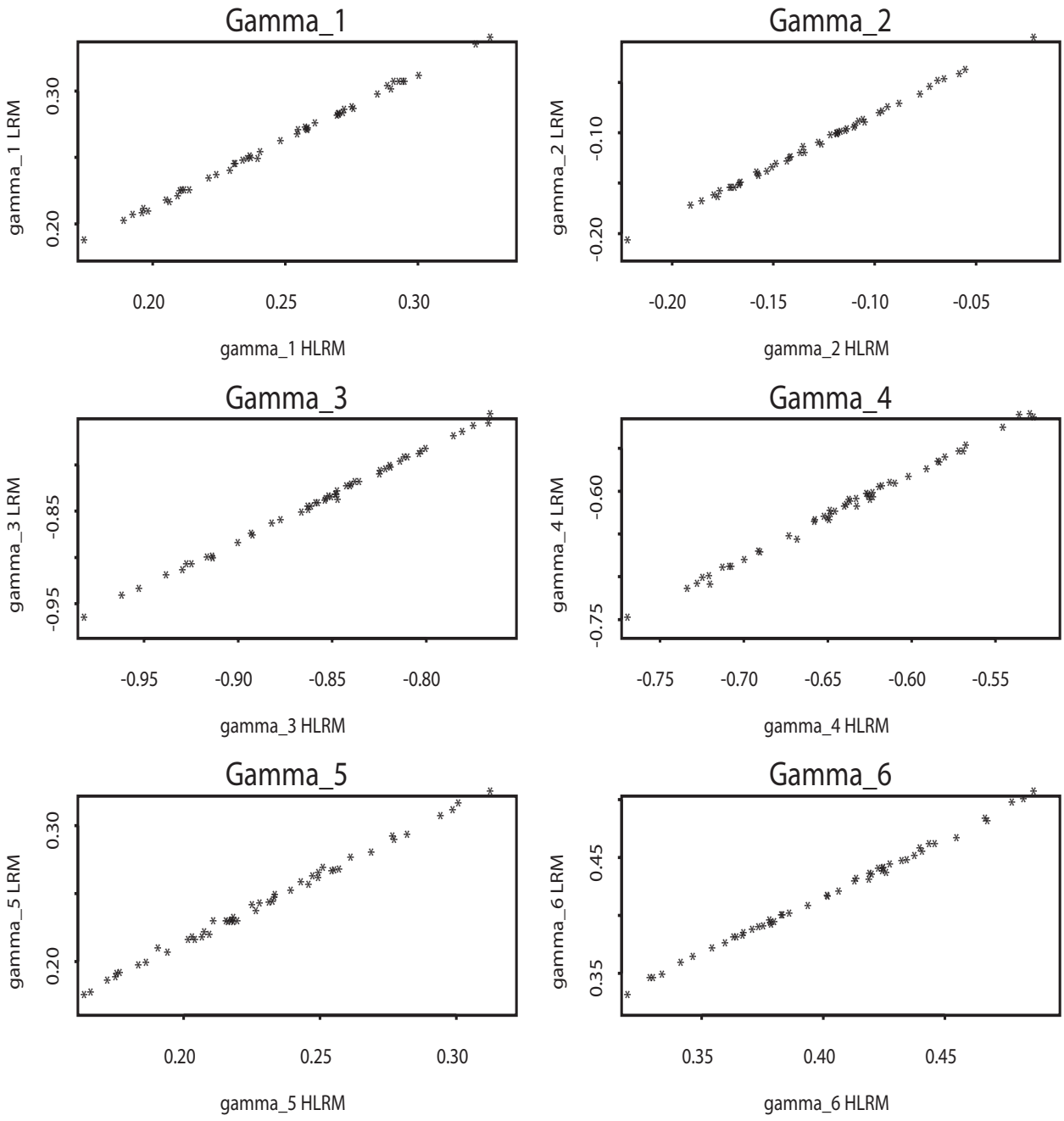

***Figure* 3 Plots of estimates of regression effect parameters over 50 replications in Condition 3.**

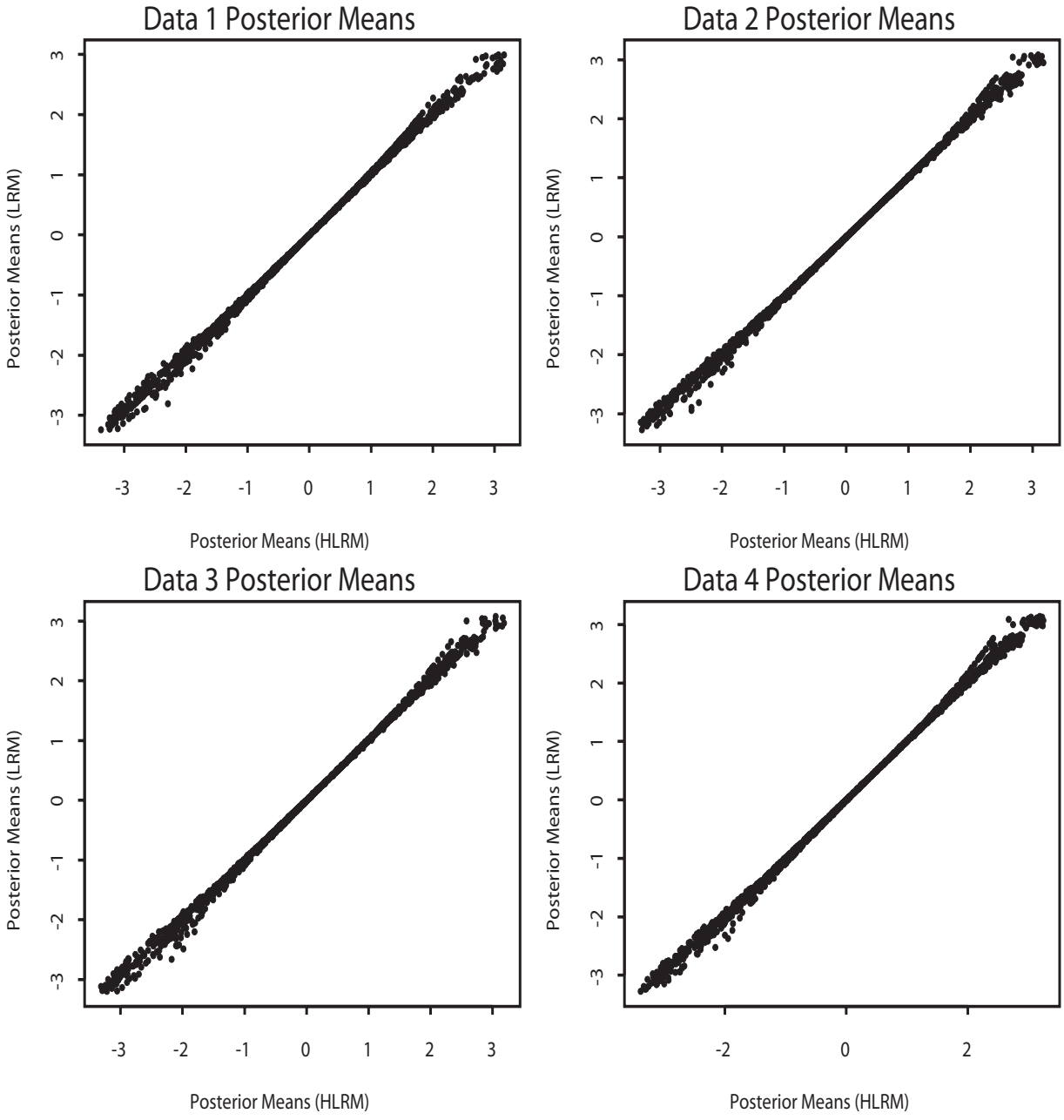

***Figure* 4 Plots of estimates of posterior means Data 1 to 4 in Condition 1.**

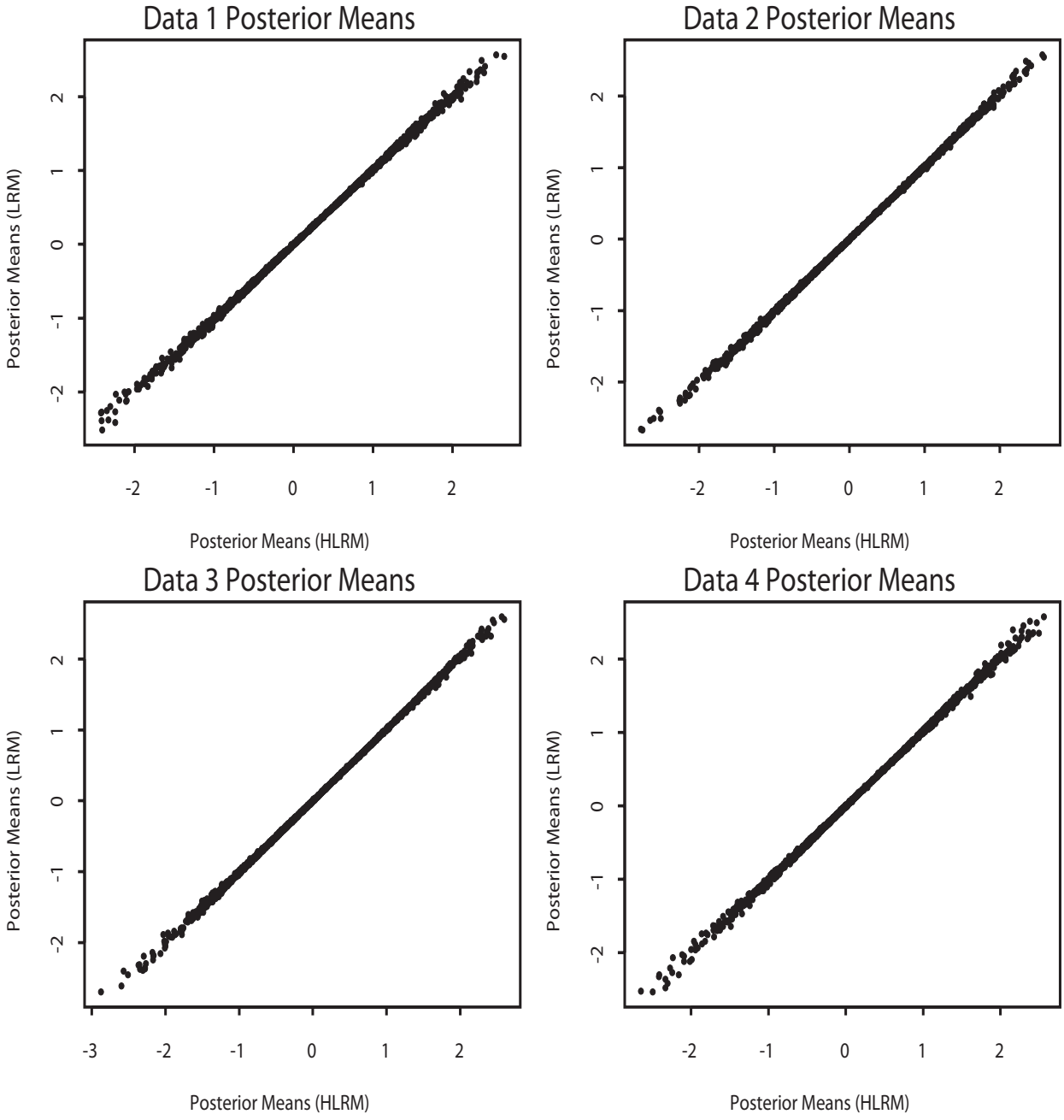

***Figure* 5 Plots of estimates of posterior means Data 1 to 4 in Condition 2.**

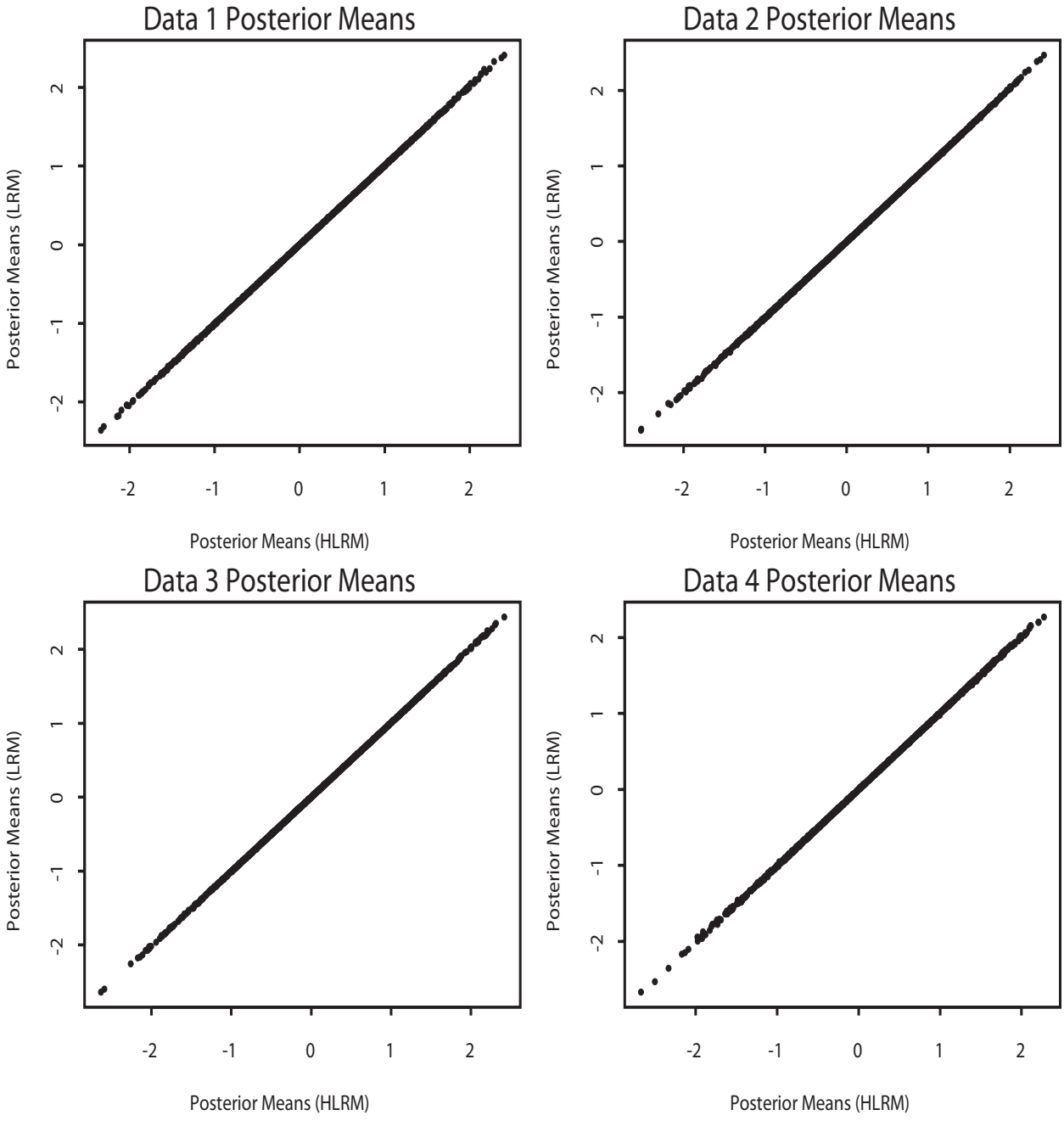

***Figure* 6 Plots of estimates of posterior means Data 1 to 4 in Condition 3.**

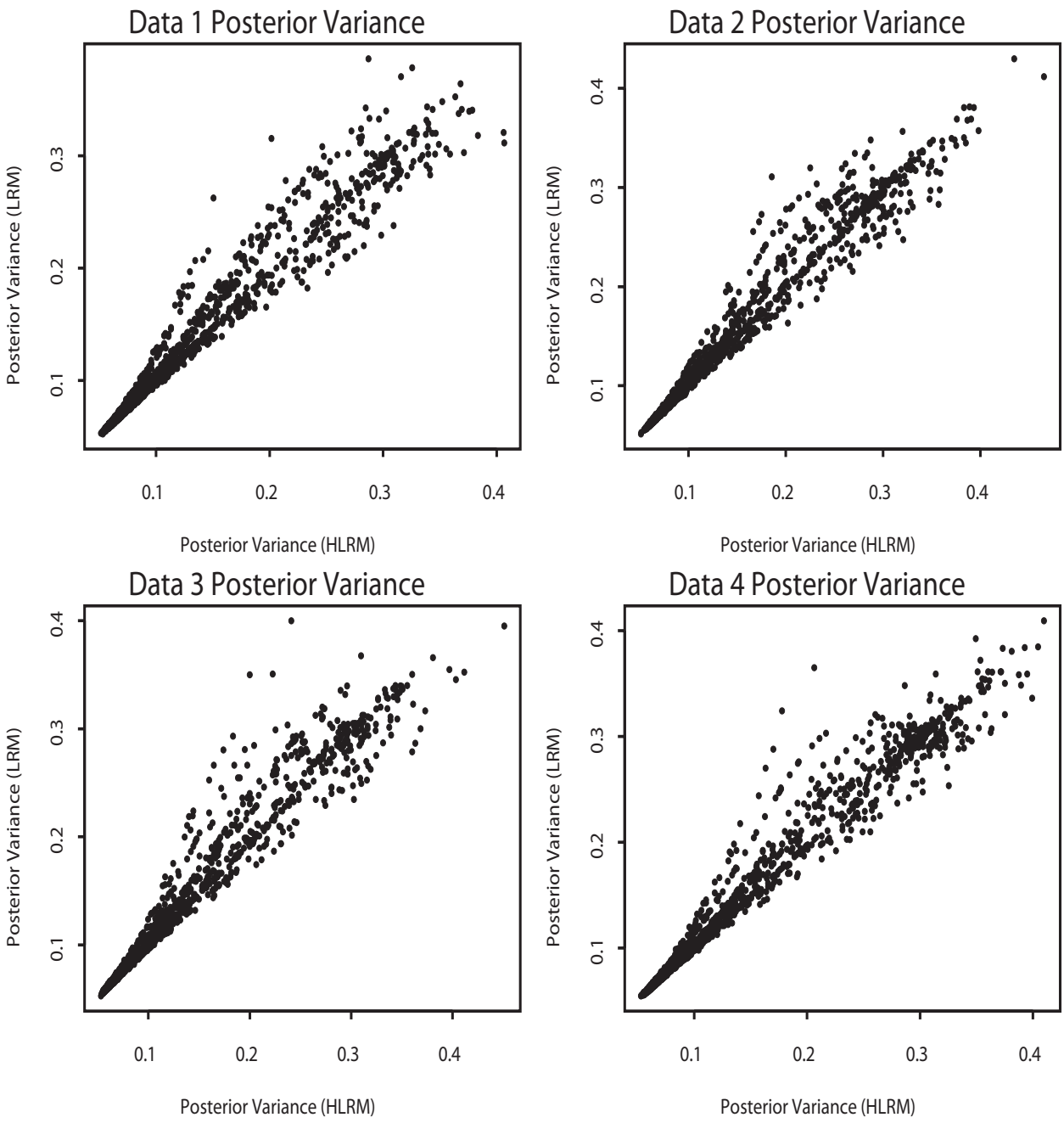

***Figure* 7 Plots of estimates of posterior variance Data 1 to 4 in Condition 1.**

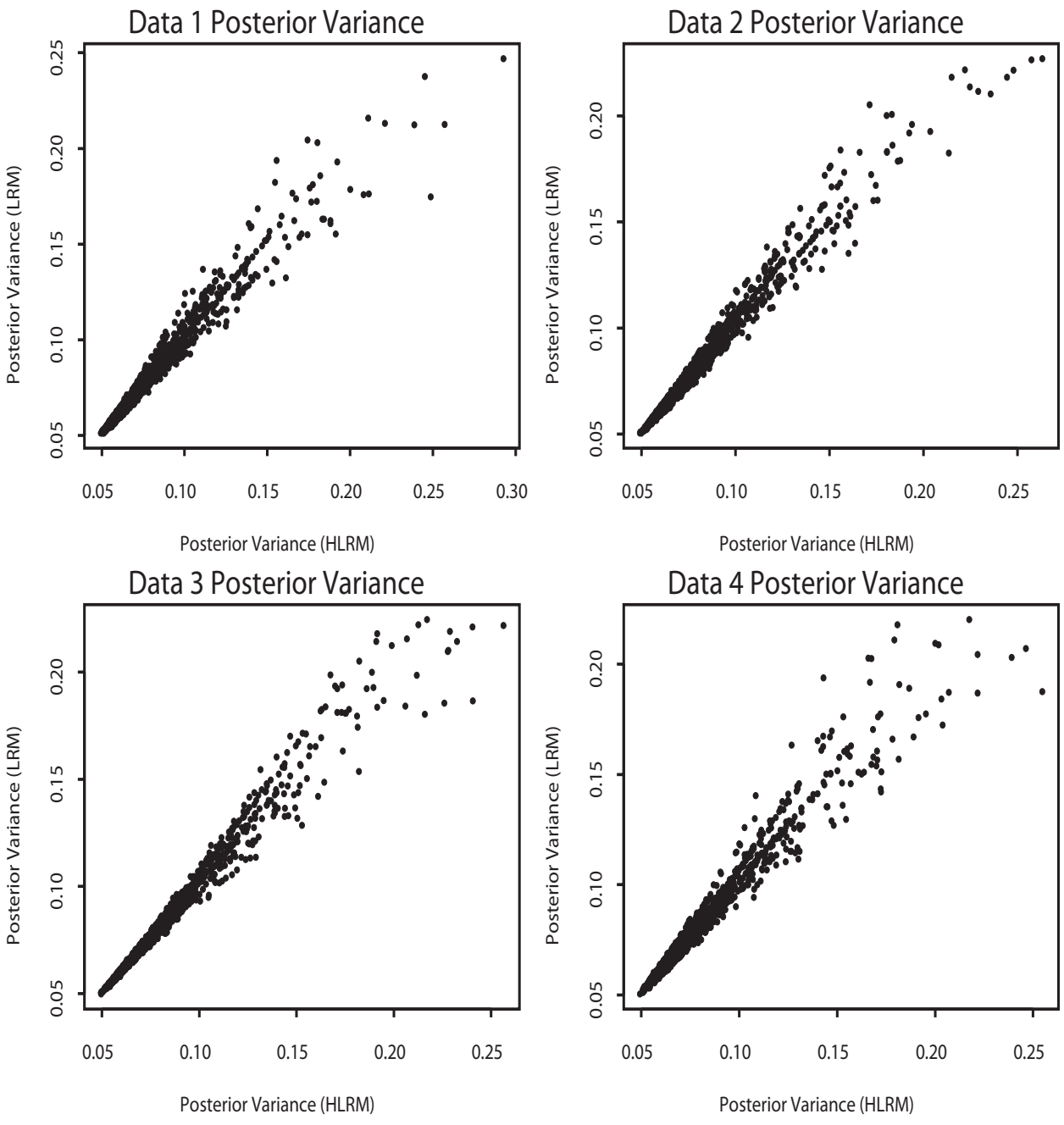

***Figure* 8 Plots of estimates of posterior variance Data 1 to 4 in Condition 2.**

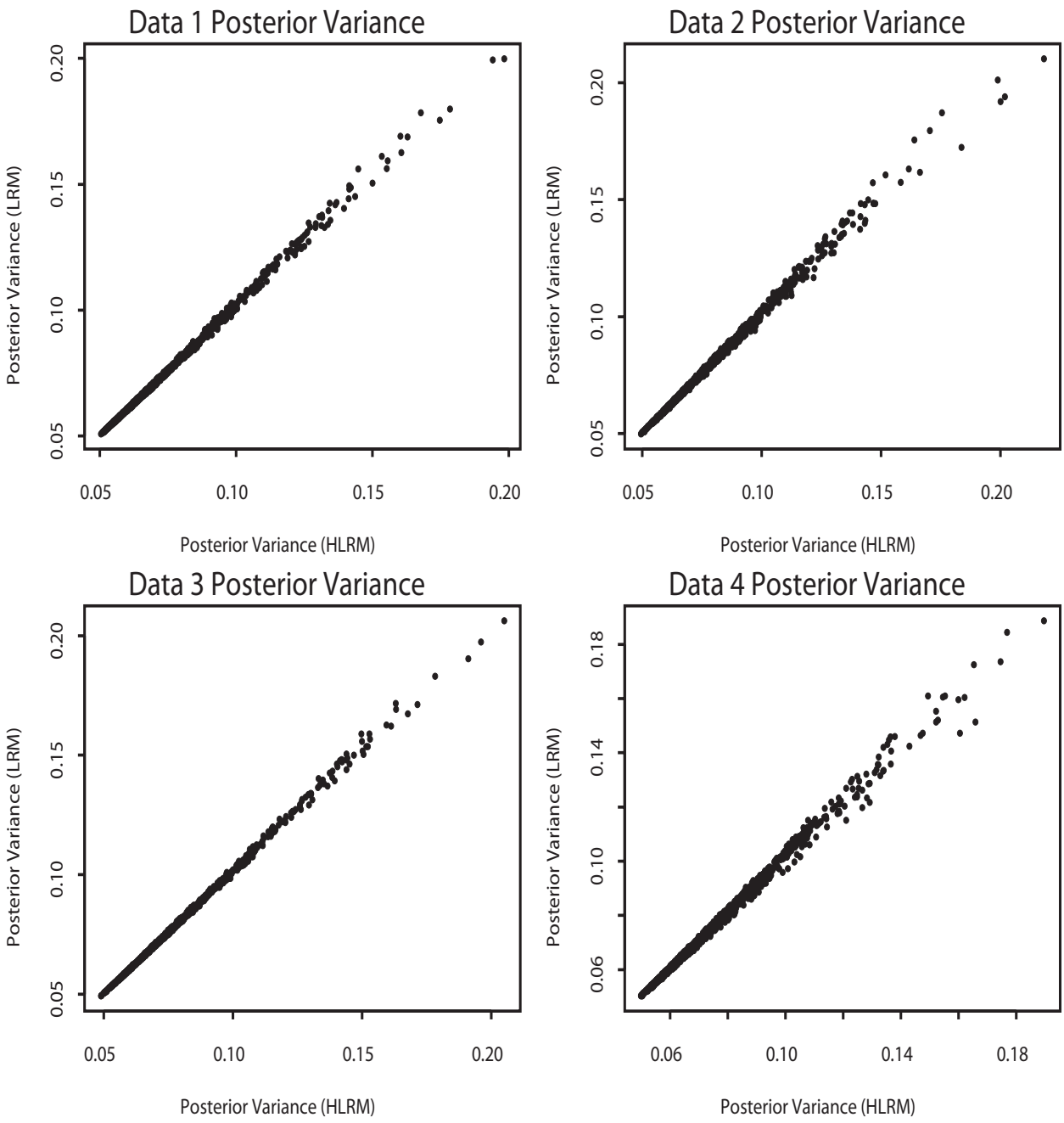

***Figure* 9 Plots of estimates of posterior variance Data 1 to 4 in Condition 3.**

The small differences between regression effect estimates across the HLRM and the LRM is also reflected in the student posterior means. In turn, this fact reflects on group mean estimates, which are a function of the posterior means for Bayes estimates and a function of the regression estimates for marginal maximum likelihood estimates. Hence, the group mean estimates between the HLRM and the LRM are quite similar. However, the posterior variance estimates for students differ substantially between the HLRM and the LRM, in particular under moderate to severe clustering.

The standard errors for the regression effects estimates between the HLRM and the LRM differ for various conditions of cluster variation. Typically, the standard errors for the regression effects are somewhat larger in the HLRM than those in the LRM. This difference is more pronounced under strong cluster variation. The simulation showed that the standard error estimates in both the HLRM and the LRM are underestimated compared to the empirical standard error. However, this underestimation reduces with increasing sample size, and the limited number of replications might not provide the most accurate assessment of empirical standard errors. It should be pointed out that, contrary to expectation, the relationship between the standard errors from the HLRM and the LRM are relatively close. This is mostly due to the fact that the residual variance is inflated in the LRM, which in turn inflates standard errors that are a function of the residual variance. Hence, if the residual variances in the LRM were accurately estimated, then the standard errors for the regression effects would be much smaller than those in Tables 3 and 4, in particular for moderate and strong clustering.

While the standard errors of the regression effects are similar between the HLRM and the LRM, the standard errors for group mean estimates are not similar; in fact, they are substantially larger under the HLRM, especially when cluster variation is moderate or strong. Because the regression-effect standard errors are similar across the models, the only difference in computing the standard errors following (21) are the off-diagonal elements of $Var(\hat{\mathbf{\Gamma}})$. Hence, a likely explanation for the differences observed in this study is that the LRM also inflates the estimates of the covariances between regression effects, resulting in underestimated group mean standard errors.

Note that in this simulation, although two overall sample sizes were used, the sample size for each school was the same across conditions. This is known as a *balanced design* in the HLM. Standard errors estimates for an unbalanced design will generally be smaller than those seen in this simulation study.

References

Allen, N., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report*(NCES-2001-509). Washington, DC: National Center for Education Statistics.

Cochran, W. G. (1977). *Sampling techniques*(3rd ed). New York: Wiley.

Cohen, J., & Jiang, T. (2002). *Direct estimation of statistics for the National Assessment of Educational Progress (NAEP).* Washington, DC: American Institutes for Research.

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, E. G., & Rust, K. F. (1992). Effective degrees of freedom for variance estimates from a complex sample survey. In *Proceedings of the Section on Survey Research Methods* (pp. 863-866). Washington, DC: American Statistical Association.

Li, D., & Oranje, A. (2006). *On the estimation of hierarchical linear models for large scale assessments* (ETS Research Rep. No. RR-06-37). Princeton, NJ: ETS.

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

Mazzeo, J., Donoghue, J., Li, D., & Johnson, M. (2005). *Marginal estimation in NAEP: Current operational procedures and AM.* Unpublished manuscript, ETS, Princeton, NJ.

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.

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

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphic Statistics, 2*(3), 309–322.

von Davier, M., & Sinharay, S. (2005). *Marginal estimation of population characteristics: Recent development and future directions.* Paper prepared for the National Center of Education Statistics.

Appendix A

Parameter Estimation for HLRM

In the univariate case, $\boldsymbol{\gamma}$ is a Q-dimensional vector of regression coefficients $(\gamma_1, \cdots, \gamma_Q)'$. Let $\boldsymbol{x}_{ij}$ be the collection (or a row) of background variables for student i in school j for scale p and $\boldsymbol{X}_j$ is a matrix of background variables of all examinees in cluster j for $j = 1, \cdots, J$, (i.e., $\boldsymbol{X}_j = [\boldsymbol{x}'_{1j}|, \cdots, |\boldsymbol{x}'_{n_j j}]'$ is an $n_j \times Q$ matrix). Then the likelihood function L for N students' responses $\boldsymbol{Y}$ to n items in a test is the total marginal likelihood, and is expressed as

$$\begin{aligned} L &= log\left[\prod_{j=1}^{J} P(\boldsymbol{y}_j|\boldsymbol{X}_j\boldsymbol{\gamma}, \boldsymbol{T}, \sigma^2)^{w_j}\right] \\ &= \sum_{j=1}^{J} w_j log\left[P(\boldsymbol{y}_j|\boldsymbol{X}_j\boldsymbol{\gamma}, \boldsymbol{T}, \sigma^2)\right] \\ &= \sum_{j=1}^{J} w_j log\left[\int P(\boldsymbol{y}_j|\boldsymbol{\theta}, \boldsymbol{u})\phi(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{X}_j\boldsymbol{\gamma}, \boldsymbol{T}, \sigma^2)d\boldsymbol{\theta}d\boldsymbol{u}\right]. \end{aligned} \tag{A1}$$

$\phi(\boldsymbol{\theta}|\boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j, \sigma^2)$ represents the conditional multivariate normal density with mean vector $\boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j$ and covariance matrix $\sigma^2\boldsymbol{I}_{n_j}$; that is, $\boldsymbol{\theta}|\boldsymbol{u}_j \sim$
$\mathcal{N}(\boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j, \sigma^2\boldsymbol{I}_{n_j})$. If the expectation of $\boldsymbol{\theta}$ is denoted by $\boldsymbol{\mu}_\theta = \boldsymbol{X}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j$, then the density function is given by

$$\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \sigma^2\boldsymbol{I}_{n_j}) = \frac{1}{(2\pi)^{\frac{n_j}{2}}|\sigma^2\boldsymbol{I}_{n_j}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)'(\sigma^2\boldsymbol{I}_{n_j})^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)\right]. \tag{A2}$$

The partial derivative of $log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \sigma^2\boldsymbol{I}_{n_j})$ with respect to $\boldsymbol{\gamma}$ is

$$\frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \sigma^2)}{\partial\boldsymbol{\gamma}} = \sigma^{-2}\boldsymbol{X}'_j(\boldsymbol{\theta} - \boldsymbol{X}_j\boldsymbol{\gamma} - \boldsymbol{X}_j\boldsymbol{u}_j). \tag{A3}$$

Therefore,

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\gamma}} &= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)} \frac{\partial \phi(\boldsymbol{\theta},\boldsymbol{u})}{\partial \boldsymbol{\gamma}} d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)} \frac{\partial \phi(\boldsymbol{\theta}|\boldsymbol{u}) p(\boldsymbol{u})}{\partial \boldsymbol{\gamma}} d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)} p(\boldsymbol{u}) \phi(\boldsymbol{\theta}|\boldsymbol{u}) \frac{\partial log \phi(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \boldsymbol{\gamma}} d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta}) \phi(\boldsymbol{\theta}) p(\boldsymbol{u}|\boldsymbol{\theta})}{P(\boldsymbol{y}_j)} \sigma^{-2} \boldsymbol{X}_j'(\boldsymbol{\theta} - \boldsymbol{X}_j\boldsymbol{\gamma} - \boldsymbol{X}_j\boldsymbol{u}_j) d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) \sigma^{-2} \boldsymbol{X}_j'(\boldsymbol{\theta} - \boldsymbol{X}_j\boldsymbol{\gamma} - \boldsymbol{X}_j\bar{\boldsymbol{u}}_j) d\boldsymbol{\theta} \\
&= \sum_{j=1}^{J} w_j \sigma^{-2} \boldsymbol{X}_j'(\tilde{\boldsymbol{\theta}} - \boldsymbol{X}_j\boldsymbol{\gamma} - \boldsymbol{X}_j\tilde{\boldsymbol{u}}_j), \qquad \text{(A4)}
\end{aligned}$$

wherein equation (A4), $\bar{\boldsymbol{u}}_j = (\bar{\boldsymbol{u}}_{j1}, \cdots, \bar{\boldsymbol{u}}_{jQ})'$, is the conditional expectation of $\boldsymbol{u}_j$ over $\phi(\boldsymbol{u}|\boldsymbol{\theta})$.

$$\bar{\boldsymbol{u}}_j = \int_u \boldsymbol{u}_j \phi(\boldsymbol{u}|\boldsymbol{\theta}) d\boldsymbol{u}, \qquad \text{(A5)}$$

and the vector form of $\tilde{\boldsymbol{u}}_j$ can be expressed as

$$\begin{aligned}
\tilde{\boldsymbol{u}}_j &= \int_{\boldsymbol{\theta}} \bar{\boldsymbol{u}}_j P(\boldsymbol{\theta}|\boldsymbol{y}_j) d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{u}} \boldsymbol{u}_j \phi(\boldsymbol{u}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{y}_j) d\boldsymbol{u} d\boldsymbol{\theta}. \qquad \text{(A6)}
\end{aligned}$$

Therefore, setting equation (A4) equal to 0 and solving for $\hat{\boldsymbol{\gamma}}$ yields

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{J} w_j \boldsymbol{X}_j' \boldsymbol{X}_j \right)^{-1} \sum_{j=1}^{J} w_j \boldsymbol{X}_j' (\tilde{\boldsymbol{\theta}}_j - \boldsymbol{X}_j \tilde{\boldsymbol{u}}_j). \qquad \text{(A7)}$$

To obtain the estimates for σ^2, it follows that

$$\begin{aligned}
\frac{\partial L}{\partial \sigma^2} &= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)} \frac{\partial \phi(\boldsymbol{\theta},\boldsymbol{u})}{\partial \sigma^2} d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)} \frac{\partial \phi(\boldsymbol{\theta}|\boldsymbol{u}) p(\boldsymbol{u})}{\partial \sigma^2} d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)} p(\boldsymbol{u}) \phi(\boldsymbol{\theta}|\boldsymbol{u}) \frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \sigma^2} d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)} \phi(\boldsymbol{\theta}) \phi(\boldsymbol{u}|\boldsymbol{\theta}) \frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \sigma^2} d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \sum_{j=1}^{J} w_j \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) \phi(\boldsymbol{u}|\boldsymbol{\theta}) \frac{\partial log\phi(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \sigma^2} d\boldsymbol{\theta} d\boldsymbol{u} \qquad \text{(A8)} \\
&= \sum_{j=1}^{J} w_j \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) P(\boldsymbol{u}|\boldsymbol{\theta}) \sum_{i=1}^{n_j} \frac{\partial log\phi(\theta_{ij}|\boldsymbol{u})}{\partial [\sigma^2]} d\boldsymbol{\theta} d\boldsymbol{u}. \qquad \text{(A9)}
\end{aligned}$$

Now it becomes more convenient to find the derivatives of $log\phi(\theta_{ij}|\mu_\theta, \sigma^2)$ with respect to σ^2:

$$\frac{\partial log\phi(\theta_{ij}|\mu_\theta,\sigma^2)}{\partial [\sigma^2]} = -\frac{1}{2}[\sigma^2]^{-1} + \frac{1}{2}[\sigma^2]^{-2}(\theta_{ij} - \boldsymbol{\gamma}'\boldsymbol{x}'_{ij} - \boldsymbol{u}'_j\boldsymbol{x}'_{ij})^2. \qquad \text{(A10)}$$

Setting the above equation equal to 0 yields:

$$\sigma^2 \sum_{j=1}^{J} w_j n_j = \sum_{j=1}^{J} w_j \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) \phi(\boldsymbol{u}|\boldsymbol{\theta}) \sum_{i=1}^{n_j} (\theta_{ij} - \mu_\theta)(\theta_{ij} - \mu_\theta) d\boldsymbol{\theta} d\boldsymbol{u}. \qquad \text{(A11)}$$

To evaluate the equation above, it is necessary to find the expression for the integration

$$\begin{aligned}
\mathcal{C} &= \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) P(\boldsymbol{u}|\boldsymbol{\theta}) \sum_{i=1}^{n_j} (\theta_{ij} - \mu_\theta)(\theta_{ij} - \mu_\theta)' d\boldsymbol{\theta} d\boldsymbol{u} \\
&= \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) P(\boldsymbol{u}|\boldsymbol{\theta}) (\boldsymbol{\theta}_j - \boldsymbol{\gamma}'\boldsymbol{X}'_j - \boldsymbol{u}'_j\boldsymbol{X}'_j)'(\boldsymbol{\theta}_j - \boldsymbol{\gamma}'\boldsymbol{X}'_j - \boldsymbol{u}'_j\boldsymbol{X}'_j) d\boldsymbol{\theta} d\boldsymbol{u}, \qquad \text{(A12)}
\end{aligned}$$

which relies on the integration to the quadratic form $(\boldsymbol{\theta}_j - \boldsymbol{\gamma}'\boldsymbol{X}'_j)'\boldsymbol{X}_j\boldsymbol{C}_j^{-1}\boldsymbol{X}'_j(\boldsymbol{\theta}_j - \boldsymbol{\gamma}'\boldsymbol{X}'_j)$ and the integration of term of $\boldsymbol{u}_j\boldsymbol{u}'_j$, which is denoted by $\boldsymbol{\mathcal{D}}$ for the time being. The integration of the later term will not be discussed here but will be included in the section discussing parameter matrix $\boldsymbol{T}$

estimation. For integrating the quadratic term, follow the theorem by Stapleton (1995, p. 51),

$$\begin{aligned}\varpi &= \int P(\boldsymbol{\theta}|\boldsymbol{y}_j)P(\boldsymbol{u}|\boldsymbol{\theta})(\boldsymbol{\theta}_j - \boldsymbol{\gamma}'\boldsymbol{X}_j')'\boldsymbol{X}_j\boldsymbol{C}_j^{-1}\boldsymbol{X}_j'(\boldsymbol{\theta}_j - \boldsymbol{\gamma}'\boldsymbol{X}_j')d\boldsymbol{\theta}d\boldsymbol{u} \\ &= (\tilde{\boldsymbol{\theta}}_j - \boldsymbol{\gamma}'\boldsymbol{X}_j')'\boldsymbol{X}_j\boldsymbol{C}_j^{-1}\boldsymbol{X}_j'(\tilde{\boldsymbol{\theta}}_j - \boldsymbol{\gamma}'\boldsymbol{X}_j') + trace(\boldsymbol{X}_j\boldsymbol{C}_j^{-1}\boldsymbol{X}_j'\tilde{\boldsymbol{\Sigma}}_j) \\ &= (\tilde{\boldsymbol{\theta}}_j - \boldsymbol{\gamma}'\boldsymbol{X}_j')'\boldsymbol{X}_j\tilde{\boldsymbol{u}}_{\boldsymbol{j}} + trace(\boldsymbol{X}_j\boldsymbol{C}_j^{-1}\boldsymbol{X}_j'\tilde{\boldsymbol{\Sigma}}_j) \\ &= \sum_{i=1}^{n_j}\left[(\tilde{\theta}_{ij} - \boldsymbol{\gamma}'\boldsymbol{x}_{ij}')'\boldsymbol{x}_{ij}\tilde{\boldsymbol{u}}_{\boldsymbol{j}}\right] + trace(\boldsymbol{X}_j\boldsymbol{C}_j^{-1}\boldsymbol{X}_j'\tilde{\boldsymbol{\Sigma}}_j). \end{aligned} \tag{A13}$$

$\tilde{\boldsymbol{\Sigma}}_j$ is a diagonal block matrix with posterior variance for each student within cluster $j = 1, \cdots, J$ as the diagonal block component. Substituting the integration (A12) gives the MML estimates for σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{J} w_j \sum_{i=1}^{n_j}\left[\tilde{\sigma}_{ij} + (\tilde{\theta}_{ij} - \boldsymbol{\gamma}'\boldsymbol{x}_{ij}')(\tilde{\theta}_{ij} - \boldsymbol{\gamma}'\boldsymbol{x}_{ij}')' + \boldsymbol{x}_{ij}\tilde{\boldsymbol{\mathcal{D}}}\boldsymbol{x}_{ij}' - 2\varpi\right]}{\sum_{j=1}^{J} w_j n_j}. \tag{A14}$$

For estimating the variance matrix $\boldsymbol{T}$ for random effects, follow the same procedure used to estimate the residual variance matrix $[\sigma^2]$:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{T}} &= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)}\frac{\partial \phi(\boldsymbol{\theta},\boldsymbol{u})}{\partial \boldsymbol{T}}d\boldsymbol{\theta}d\boldsymbol{u} \\ &= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)}\frac{\partial \phi(\boldsymbol{\theta}|\boldsymbol{u})p(\boldsymbol{u})}{\partial \boldsymbol{T}}d\boldsymbol{\theta}d\boldsymbol{u} \\ &= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j|\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)}p(\boldsymbol{u})\phi(\boldsymbol{\theta}|\boldsymbol{u})\frac{\partial logP(\boldsymbol{u})}{\partial \boldsymbol{T}}d\boldsymbol{\theta}d\boldsymbol{u} \\ &= \sum_{j=1}^{J} w_j \int \frac{P(\boldsymbol{y}_j,\boldsymbol{\theta},\boldsymbol{u})}{P(\boldsymbol{y}_j)}\frac{\partial logP(\boldsymbol{u})}{\partial \boldsymbol{T}}d\boldsymbol{\theta}d\boldsymbol{u} \\ &= \sum_{j=1}^{J} w_j \int P(\boldsymbol{\theta}|\boldsymbol{y}_j)P(\boldsymbol{u}|\boldsymbol{\theta})\frac{\partial logP(\boldsymbol{u})}{\partial \boldsymbol{T}}d\boldsymbol{\theta}d\boldsymbol{u}. \end{aligned} \tag{A15}$$

The derivative of $logP(\boldsymbol{u})$ with respect to $\boldsymbol{T}$ can be simplified as:

$$\begin{aligned}\frac{\partial logP(\boldsymbol{u})}{\partial \boldsymbol{T}} &= \frac{1}{2}diag\left[\boldsymbol{T}^{-1} - \boldsymbol{T}^{-1}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{T}^{-1}\right] - \left[\boldsymbol{T}^{-1} - \boldsymbol{T}^{-1}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{T}^{-1}\right] \\ &= \frac{1}{2}diag\left[\boldsymbol{T}^{-1}(\boldsymbol{T} - \boldsymbol{u}\boldsymbol{u}')\boldsymbol{T}^{-1}\right] - \left[\boldsymbol{T}^{-1}(\boldsymbol{T} - \boldsymbol{u}\boldsymbol{u}')\boldsymbol{T}^{-1}\right]. \end{aligned} \tag{A16}$$

Hence,

$$\frac{\partial L}{\partial \boldsymbol{T}} = \frac{1}{2}\sum_{j=1}^{J} w_j\left(diag\left[\boldsymbol{T}^{-1}(\boldsymbol{T} - \boldsymbol{\mathcal{E}})\boldsymbol{T}^{-1}\right] - \left[\boldsymbol{T}^{-1}(\boldsymbol{T} - \boldsymbol{\mathcal{E}})\boldsymbol{T}^{-1}\right]\right). \tag{A17}$$

where $\boldsymbol{\mathcal{E}}$ indicates the following integration:

$$\boldsymbol{\mathcal{E}} = \frac{1}{\sum_{j=1}^{J} w_j} \sum_{j=1}^{J} w_j \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) P(\boldsymbol{u}|\boldsymbol{\theta}) \boldsymbol{u}\boldsymbol{u}' d\boldsymbol{\theta} d\boldsymbol{u}, \tag{A18}$$

This is similar to A12 and can be written as:

$$\begin{aligned} \boldsymbol{\mathcal{E}} \sum_{j=1}^{J} w_j &= \sum_{j=1}^{j} w_j \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) P(\boldsymbol{u}|\boldsymbol{\theta}) \boldsymbol{u}\boldsymbol{u}' d\boldsymbol{\theta} d\boldsymbol{u} \\ &= \sum_{j=1}^{j} w_j \left[Var(\boldsymbol{u}|\boldsymbol{\theta}) + \int P(\boldsymbol{\theta}|\boldsymbol{y}_j) \bar{\boldsymbol{u}}\bar{\boldsymbol{u}}' d\boldsymbol{\theta} d\boldsymbol{u} \right] \\ &= \sum_{j=1}^{j} w_j \left[Var(\boldsymbol{u}|\boldsymbol{\theta}) + \tilde{\boldsymbol{u}}_j \tilde{\boldsymbol{u}}_j' + \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}_j' \tilde{\boldsymbol{\Sigma}}_j \boldsymbol{X}_j \boldsymbol{C}_{jt}^{-1} \right], \end{aligned} \tag{A19}$$

where $\bar{\boldsymbol{u}}$ in A21 stands for the conditional expectation of $\boldsymbol{u}_j$ given students' abilities in cluster j (i.e. $\boldsymbol{\theta}_j$):

$$\bar{\boldsymbol{u}} = \int \boldsymbol{u}_j P(\boldsymbol{u}_j|\boldsymbol{\theta}_j) d\boldsymbol{u}_j. \tag{A20}$$

For one subscale case,

$$\boldsymbol{\mathcal{E}} \sum_{j=1}^{J} w_j = \sum_{j=1}^{j} w_j \left[\boldsymbol{C}_{jt}^{-1} \sigma^2 + \tilde{\boldsymbol{u}}_j \tilde{\boldsymbol{u}}_j' + \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}_j' \left[\tilde{\sigma}_{ij}^2 \right] \boldsymbol{X}_j \boldsymbol{C}_{jt}^{-1} \right]. \tag{A21}$$

Thus, $\boldsymbol{T}$ can be estimated by setting A17 equal to zero and obviously having $\boldsymbol{T} = \boldsymbol{\mathcal{E}}$,

$$\hat{\boldsymbol{T}} = \frac{1}{\sum_{j=1}^{J} w_j} \sum_{j=1}^{j} w_j \left[\boldsymbol{C}_{jt}^{-1} \sigma^2 + \tilde{\boldsymbol{u}}_j \tilde{\boldsymbol{u}}_j' + \boldsymbol{C}_{jt}^{-1} \boldsymbol{X}_j' \left[\tilde{\sigma}_{ij}^2 \right] \boldsymbol{X}_j \boldsymbol{C}_{jt}^{-1} \right]. \tag{A22}$$

Appendix B

Extension to More General Two-Level Hierarchical Linear Models

For one subscale case, suppose that the Level 2 model is conditional on a set of school/cluster characteristic variables, $\boldsymbol{V}_j$. Then the model is:

$$\boldsymbol{\gamma}_j \quad = \quad \boldsymbol{V}_j\boldsymbol{\gamma} + \boldsymbol{u}_j, \tag{B1}$$

and the combined model becomes

$$\boldsymbol{\theta}_{jt} \quad = \quad \boldsymbol{X}_j\boldsymbol{V}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j + \boldsymbol{\varepsilon}_j. \tag{B2}$$

In general, Raudenbush and Bryk (2002) suggested writing the two-level HLM as:

$$\boldsymbol{\theta}_{jt} \quad = \quad \boldsymbol{A}_{fj}\boldsymbol{\gamma} + \boldsymbol{A}_{rj}\boldsymbol{u}_j + \boldsymbol{\varepsilon}_j. \tag{B3}$$

By following the same procedure developed above, it is possible to obtain parameter estimates for a general two-level HLRM.

Appendix C

Assumptions

C1: IRT local dependence assumption.

C2: The assumption of Level 1 random variable $\varepsilon_{ij} \sim i.i.d.N(0, \sigma^2)$; Level 2 random variable $\boldsymbol{u}_j \sim i.i.d.N(\mathbf{0}, \boldsymbol{T})$; and ε_{ij} is independent with $\boldsymbol{u}_j$ for $i = 1, \cdots, n_j, j = 1, \cdots, J$.

C3: The abilities of students within the same clusters are not necessarily independent; rather, the abilities of students within the same clusters are independent if the cluster effects are controlled. If the assumption on **A2** holds, it follows naturally that $\phi(\boldsymbol{\theta}_{ij}|\boldsymbol{u}_j) \sim i.i.d.N(\boldsymbol{\gamma}'\boldsymbol{x}'_{ij} + \boldsymbol{u}'_j\boldsymbol{x}'_{ij}, \sigma^2)$. Obviously, $\phi(\boldsymbol{\theta}_j|\boldsymbol{u}_j)$ can be written as the product of $\phi(\theta_{ij}|\boldsymbol{u}_j)$ for abilities θ_{ij} of students within the same cluster. Intuitively, this assumption makes sense. After cluster effects are accounted for, student abilities can be considered independent of each other.

Appendix D

Alternative Numerical Integration

The univariate estimation methodology can easily be extended to the multivariate case, which is detailed in Li and Oranje (2006). The posterior mean $\tilde{\boldsymbol{\theta}}_i$ and posterior variance $\tilde{\boldsymbol{\Sigma}}_j$ for $j = 1, \cdots, J$ can be obtained through a multivariate numerical quadrature. A relatively efficient alternative is a procedure for asymptotic corrections of multivariate posterior moments using a factored likelihood function (Thomas, 1993). Let $h(\boldsymbol{\theta}_j)$ be proportional to the posterior distribution of $\boldsymbol{\theta}_j$; that is,

$$h(\boldsymbol{\theta}_j) \quad = \quad -log\left[f(\boldsymbol{y}_i|\boldsymbol{\theta})\phi(\boldsymbol{\theta})\right], \tag{D1}$$

where

$$logf(\boldsymbol{y}_j|\boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{n_j}\sum_{i=1}^{n}\left[y_{ik}logP_k(\theta_{ik}) + (1-y_{ik})log(1-P_k(\theta_{ik}))\right], \tag{D2}$$

with $P_k(\theta_{ik})$ a specified item response model (e.g., the 3PL model). The joint distribution for $\boldsymbol{\theta}$ is

$$\phi(\boldsymbol{\theta}) \quad = \quad \frac{1}{(2\pi)^{\frac{n_j}{2}}|\triangle_j|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{\theta}_j-\boldsymbol{X}_j\boldsymbol{\gamma})'\triangle_j^{-1}(\boldsymbol{\theta}_j-\boldsymbol{X}_j\boldsymbol{\gamma})}. \tag{D3}$$

The covariance matrix in (D3), $\triangle_j$, is given by

$$\triangle_j \quad = \quad \boldsymbol{X}_j\boldsymbol{T}\boldsymbol{X}_j' + \sigma^2\boldsymbol{I}_{n_j}. \tag{D4}$$

The first step to maximize $h(\boldsymbol{\theta}_j)$ is to find the partial derivative of $h(\boldsymbol{\theta}_j)$ with respect to $\boldsymbol{\theta}_j$, which can be expressed as

$$\frac{\partial h(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \quad = \quad -diag\left(h_1(\theta_1), \cdots, h_k(\theta_k), \cdots, h_{n_j}(\theta_{n_j})\right) + \triangle_j^{-1}(\boldsymbol{\theta}_j - \boldsymbol{X}_j\boldsymbol{\gamma}). \tag{D5}$$

Due to the factored likelihood function in (D2), the first part of the partial derivative in (D5) is a diagonal matrix with diagonal elements

$$h_k(\theta_k) \quad = \quad \sum_{i=1}^{n}\left(\frac{y_{ik}}{P_i(\theta_{ik})} - \frac{1-y_{ik}}{1-P_i(\theta_{ik})}\right)P_i'(\theta_{ik}). \tag{D6}$$

The second derivative of $h(\boldsymbol{\theta}_j)$ with respect to $\boldsymbol{\theta}_j$ can be further simplified and expressed as

$$\frac{\partial h^2(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j^2} \quad = \quad -diag\left(\frac{\partial h_1(\theta_1)}{\partial \theta_1}, \cdots, \frac{\partial h_k(\theta_k)}{\partial \theta_k}, \cdots, \frac{\partial h_{n_j}(\theta_{n_j})}{\partial \theta_{n_j}}\right) + \triangle_j^{-1}. \tag{D7}$$

The diagonal part of the second derivative in (D7) is

$$\frac{\partial h_k(\theta_k)}{\partial \theta_k} = \sum_i^n \left(\frac{y_{ik}(Da_i)^2 c_i e^{DL}}{(c_i + e^{DL})^2} - \frac{Da_i}{1 - c_i} P_i'(\theta_{ik}) \right), \tag{D8}$$

where D is the scaling factor in the 3PL model, $L = a_i(\theta_{ik} - b_i)$, and $P_i'(\theta_{ik})$ is the derivative of $P_i(\theta_{ik})$ with respect to θ_k; that is,

$$P_i'(\theta_{ik}) = \frac{\partial P_i(\theta_{ik})}{\partial \theta_k} = \frac{Da_i}{1 - c_i}(1 - P_i(\theta_{ik}))(P_i(\theta_{ik}) - c_i). \tag{D9}$$

The Newton-Raphson or Fisher scoring algorithm can be used to find the posterior mode $\tilde{\boldsymbol{\theta}}_j$ and covariance $\tilde{\boldsymbol{\Sigma}}_j$ evaluated at this mode. Further asymptotic corrections follow similar to Thomas (1993).