Appendix A: Formulation for the Four-Level Two-Order HIRT Model

As stated in the main document, the ML-HIRT model is not necessarily limited to three levels; more than three levels are possible in real testing situations. For example, students may be sampled from specific groups (e.g., classes), and those in the same class would share the same teacher and curriculum, so the measurements in the same class can no longer considered to be independent. In this case, the students are nested within classes, and a between-class model (Level 4) can be formulated by treating the regression coefficients at Level 3 as random variables for classes. Accordingly, the Level-2 and Level-3 models can be rewritten as

$$\theta^{(2)}_{nct} = \boldsymbol{\omega}_{nct}\boldsymbol{\beta}_{nc} + \psi_{nct} \tag{A1}$$

with

$$\psi_{nct} \sim N(0, \sigma^2_{\psi(ct)}) \tag{A2}$$

and

$$\beta_{ncd} = \boldsymbol{\kappa}_{nc}\boldsymbol{\gamma}_{cd} + \varsigma_{ncd}\,, \tag{A3}$$

with

$$\boldsymbol{\varsigma}_{nc} \equiv [\varsigma_{nc0}, \varsigma_{nc1}, \ldots \varsigma_{nch}] \sim N(0, \boldsymbol{\Sigma}_{\varsigma(c)})\,, \tag{A4}$$

respectively, where the subscript *c* is incorporated in the formulations to represent the random effects associated with different classes of students and to indicate the predictors that depend not only on person *n* but also on class *c*, and the definitions of the other parameters are the same as in the main document. As a result, the Level-4 model can be expressed as

$$\gamma_{cdk} = \xi_c \boldsymbol{\varphi}_{dk} + \nu_{cdk}\,, \text{ for } \; k = 0,1,\cdots,m \tag{A5}$$

with

$$\boldsymbol{\nu}_{cd} \equiv [\nu_{cd0}, \nu_{cd1}, \ldots \nu_{cdm}] \sim N(0, \boldsymbol{\Sigma}_{\nu(d)})\,, \tag{A6}$$

where ξ_c is a set of observed class-level variables for class *c*, φ_{dk} is the vector of class-level regression coefficients for the *dk*th coefficient at Level 3, and $\boldsymbol{\nu}_{cd}$ is the class-level regression residual vector for the *d*th growth factor, which is assumed to be mutually independent of the other levels' residuals. The estimation of the four-level model is straightforward because no further constraints need to be imposed on these equations; however, it becomes computationally burdensome because many parameter values need to be estimated.

Appendix B: Simulation Process for the ML-HIRT Model

The generated values resulted from an analysis of the empirical example consisting of dichotomous items, using a linear growth approach for the purposes of demonstration. The simulation process consisted of the following steps. First, the random intercept (β_0) and random slope (β_1) were generated by incorporating the Level-3 predictor (i.e., region; urban regions were coded as 1 and rural regions were coded as -1) and by assuming that the Level-3 residuals follow a multivariate normal distribution, using the true values for Level 3, together with Equations 5 and 6. More specifically, the random intercept and random slope for the set of examinees from urban regions can be expressed as $\beta_{n0(urban)} = 0.190 + \varsigma_{n0(urban)}$ and $\beta_{n1(urban)} = 0.396 + 0.032 + \varsigma_{n1(urban)}$, respectively. The random intercept and random slope for the set of examinees from rural regions can be expressed as $\beta_{n0(rural)} = -0.190 + \varsigma_{n0(rural)}$ and $\beta_{n1(rural)} = 0.396 - 0.032 + \varsigma_{n1(rural)}$, respectively. The Level-3 residual variances were assumed to be homogenous for the two regions, and the residuals were generated from a mean vector of zero and a variance–covariance matrix of $\begin{bmatrix} 0.428 & \\ 0.073 & 0.039 \end{bmatrix}$. Second, a linear growth model was used to regress the second-order latent traits against the time-based predictors with random regression coefficients from Level 3 across the four occasions, using the true values related to Level 2, together with Equations 3 and 4. Finally, the first-order latent traits for each occasion were specified through the combination of the second-order latent trait with the corresponding weights of the first-order traits (i.e., factor loadings) and Level-1 residuals, using Equation 2. The item responses of the examinees were generated according to Equation 1 once the latent traits and item parameters had been determined.

Appendix C: Prior Distribution and Convergence Monitoring

Before implementing Bayesian estimation, a prior distribution must be specified for each parameter in the ML-HIRT model. In both simulations, normal prior distributions with means of 0 and variances of 4 were assumed for the location and regression parameters, log-normal prior distributions with means of 0 and variances of 1 were assumed for the slope parameters, a beta prior distribution with both hyperparameters equal to 1 was assumed for the pseudo-guessing parameter, normal prior distributions with means of 0.5 and variances of 10 were assumed for the factor loadings, a gamma prior distribution with both hyperparameters equal to 0.01 was assumed for the inverse of the residual variances, and a Wishart distribution with a diagonal scale matrix and two degrees of freedom was assumed for the inverse variance–covariance matrix.

The multivariate potential scale reduction factor (Brooks & Gelman, 1998) was used to determine the number of iterations, assuming three parallel chains for five randomly selected simulated data sets for all of the analytical models. The results showed that 15,000 iterations were sufficient to reach stationarity, with the first 5,000 iterations defined as the burn-in because all of the multivariate potential scale reduction factors were close to 1.2. In addition, the Geweke convergence diagnostic (Geweke, 1992) was used to compare the mean of the parameter's posterior distribution from the first 10% of the chain to that from the last 50% of the chain. Additionally, a *Z* statistic was calculated for statistical hypothesis testing, and the results showed that all *Z* statistics did not fall in the critical region, suggesting that no evidence was found against convergence. The history plots of the generated sequences displayed good convergence: the three chains mixed very well, and there was no change point or trend in the plot. The trace plots indicated that all of the estimated parameters became stationary at 15,000 iterations.

Appendix D: Parameter Recovery for Different Times and Different Item Types

The item parameter recoveries for different times were compared to assess the estimation accuracy for common and unique items on different occasions. As shown in Table D1, for the large sample of 4,000 examinees, the RMSE values ranged from 0.033 to 0.461 (*M* = 0.127) for the difficulty parameters and from 0.019 to 0.275 (*M* = 0.095) for the slope parameters for occasion 1, from 0.029 to 0.263 (*M* = 0.088) for the difficulty parameters and from 0.024 to 0.215 (*M* = 0.095) for the slope parameters for occasion 2, from 0.033 to 0.286 (*M* = 0.118) for the difficulty parameters and from 0.019 to 0.215 (*M* = 0.081) for the slope parameters for occasion 3, and from 0.043 to 0.337 (*M* = 0.139) for the difficulty parameters and from 0.019 to 0.215 (*M* = 0.085) for the slope parameters for occasion 4. The differences in parameter recovery for the different occasions were negligible. It can therefore be concluded that the item parameter estimation was nearly independent of the occasion. Similar results were obtained for the small sample of 1,000 examinees, although the RMSE values were slightly larger for the small sample.

The lower panel of Table D1 shows the differences in RMSE between the unique and common items across occasions. When the sample size was 4,000, the unique items were found to have slightly larger mean RMSE values (0.142 and 0.108 for the difficulty and slope parameters, respectively) than the common items (0.104 and

0.079 for the difficulty and slope parameters, respectively), indicating that the common items could be estimated more precisely than the unique items. The same conclusions were obtained for the sample size of 1,000, although the parameter estimation was inferior to that for the larger sample.

Table D1. RMSE of Item Parameter Estimates for Different Administration Occasions and Different Item Types in the 3P-ML-HIRT Model

				Sample Size			
		True Values		4,000		1,000	
Parameter		Difficulty	Slope	Difficulty	Slope	Difficulty	Slope
Administration							
Occasion 1	Mean	-0.580	1.057	0.127	0.095	0.243	0.163
	SD	1.473	0.446	0.096	0.048	0.162	0.081
	Max	2.791	2.210	0.461	0.275	0.658	0.363
	Min	-4.224	0.244	0.033	0.019	0.056	0.031
Occasion 2	Mean	0.295	1.039	0.088	0.095	0.170	0.167
	SD	0.743	0.431	0.053	0.051	0.108	0.092
	Max	1.547	2.371	0.263	0.215	0.493	0.370
	Min	-2.133	0.410	0.029	0.024	0.056	0.031
Occasion 3	Mean	0.619	0.961	0.118	0.081	0.242	0.137
	SD	1.192	0.426	0.080	0.045	0.166	0.072
	Max	2.928	2.059	0.286	0.215	0.701	0.359
	Min	-1.955	0.244	0.033	0.019	0.056	0.040
Occasion 4	Mean	0.989	0.918	0.139	0.085	0.262	0.138
	SD	1.078	0.407	0.081	0.045	0.167	0.072
	Max	3.367	2.059	0.337	0.215	0.670	0.359
	Min	-1.955	0.244	0.043	0.019	0.056	0.031
Item Type							
Common	Mean	0.679	0.948	0.104	0.079	0.199	0.135
	SD	0.988	0.393	0.074	0.440	0.147	0.715
	Max	2.928	2.059	0.283	0.215	0.620	0.359
	Min	-2.367	0.244	0.033	0.019	0.056	0.031
Unique	Mean	-0.472	1.085	0.142	0.108	0.279	0.183
	SD	1.456	0.472	0.090	0.502	0.160	0.900
	Max	3.367	2.371	0.461	0.275	0.701	0.370
	Min	-4.224	0.352	0.029	0.041	0.060	0.057

Note: RMSE = root mean square error.

Appendix E: A Simulation with the Incorporation of Extreme Item Difficulty

To test the hypothesis that the ranges of the item parameter values affect the estimation precision for the examinees, a simulation was performed in which a set of higher and lower item difficulty parameters was incorporated on the fourth testing occasion for the three tests, and the large sample size was used. Specifically, the same simulation design as described above in the large sample size was used, but four items with relatively low and high difficulty were added to each of the three tests for the fourth occasion. The difficulty parameter values were set to -5, -4, 4, and 5, respectively, for the four items. The values of all of the discrimination parameters were set to one because the mean discrimination parameter was nearly equal to unity. The values of all of the pseudo-guessing parameters were set to 0.147. Such extreme difficulty parameter values are not rarely observed in real testing situations. Table E1 shows the person parameter recovery results for the four occasions in terms of the mean RMSE across replications for the second- and first-order latent traits. As expected, the mean RMSE decreased to 0.432, 0.557, and 0.700 for the three first-order latent traits and to 0.353 for the second-order latent trait for the fourth occasion. A much wider range of item difficulty parameters (for example, between -6 and 9) is expected to provide more accurate latent trait estimation because the true latent traits were found between -5.40 and 8.94 on the fourth occasion.

Table E1. Mean RMSE of Person Parameter Estimates for the 3PL-ML-HIRT Model

Condition	Occasion 1	Occasion 2	Occasion 3	Occasion 4
1st-Order Latent Trait in Test 1	0.270	0.352	0.429	0.432
1st-Order Latent Trait in Test 2	0.496	0.473	0.544	0.557
1st-Order Latent Trait in Test 3	0.414	0.413	0.552	0.700
2nd-Order Latent Trait	0.246	0.245	0.297	0.353

Notes: Higher and lower item difficulty parameters were incorporated for the fourth measurement occasion.

Appendix F: Two Empirical Examples

Example 1: Basic Ability Assessment

As an example with dichotomous items, the Basic Ability Assessment was administered in multiple-choice format to 4,007 junior high school students (2,418 urban students and 1,589 rural students) from the seventh to the twelfth grades in four studies conducted in Taiwan (Chang, 2007). Three tests measured the students' basic abilities, including analytical ability, reading ability, and mathematical ability, and these three tests can be considered to measure the first-order latent traits governed by a second-order latent trait describing overall basic ability. Because the items in the three tests were partially replaced by a new set of items as time progressed, the tests did not consist of exactly the same items for each measurement; instead, minimum requirements for the number of common items were satisfied, and the inclusion of more anchor items led to more precise parameter estimation. The test measuring analytical ability consisted of 18, 13, 13, and 9 items administered at four different times; the test measuring reading ability consisted of 10, 10, 6, and 7 items administered at four different times; and the test measuring mathematical ability consisted of 20, 19, 17, and 19 items administered on four separate occasions. There were 12 common items between occasions 1 and 2 for the three tests; 12 common items between occasions 2 and 3 for the three tests; 15 common items between occasions 3 and 4 for the three tests; and a total of 12 common items for all three occasions for the three tests. The region in which the students resided was treated as the regression predictor at Level 3 for specifying the variation in the growth trajectories.

Four questions were of particular interest in the model comparison in this study. First, do the items share a common slope parameter, and is the pseudo-guessing parameter necessary? To answer this question, the 1P-, 2P-, and 3P-ML-HIRT models (with linear growth) were fit to the data, and the fit of the model to the data was assessed using the Bayesian DIC. A smaller DIC value indicates a better fit of the model to the data. Second, is a higher-order structure necessary to account for the relationship between latent traits? To answer this question, the ML-HIRT model was compared with the ML-IRT model (i.e., all the test items measured a single latent trait) when fitting to the data set. Third, is a linear latent growth model sufficient to yield a good fit of the model to the data, or is a nonlinear (quadratic) growth model required to fit the data? To answer this question, two types of growth models, with linear and nonlinear growth, were fit and compared. Fourth, did the urban and rural students have systematically different growth trajectories? To answer this question, the ML-HIRT model with Level-3 predictors was compared with the same model without Level-3 predictors.

With respect to the first question, the DIC value was 690,366 for the 1P-ML-HIRT model, 680,933 for the 2P-ML-HIRT model, and 680,044 for the 3P-ML-HIRT model. The 3P-ML-HIRT model with linear growth therefore yielded a superior fit because of its smaller DIC. With respect to the second question, when linear growth was considered, the 3P-ML-HIRT model yielded a better fit (DIC = 680,044) than the 3P-ML-IRT model (DIC = 683,034), and a higher-order structure of latent traits was therefore necessary and was not neglected. To answer the third question, the 3P-ML-HIRT model with quadratic growth was compared with the same model with linear growth, and the results showed that the quadratic growth model yielded a better fit (DIC = 679,483). Finally, the quadratic-growth 3P-ML-HIRT model with Level-3 predictors was compared with the same model without Level-3 predictors, and the results showed that the former yielded a better fit (as indicated by its smaller DIC value of 679,329). Because the Level-3 regression coefficients of the linear growth factor (β_1) and quadratic growth factor (β_2) were very small and were not significantly different from zero, only the variation in the random intercept (β_0) was considered for the urban and rural students. The parameter estimation is summarized as follows. The estimates were between -4.65 and 3.18 (M = 0.46) for the difficulty parameters and between 0.28 and 2.60 (M = 1.07) for the slope parameters, and the estimate was 0.16 for the pseudo-guessing parameter. The factor loadings were 1.00, 1.06, and 1.38 for analytical ability, reading ability, and mathematical ability, respectively. The grand mean of the random intercept (β_0) was 0.21 for the urban students and -0.21 for the rural students, indicating that the examinees in urban and rural regions exhibited different initial status conditions and that the growth slopes were not significantly affected by the type of region (urban versus rural).

Finally, to verify the fit of the best-fitting quadratic growth 3P-ML-HIRT model with Level-3 predictors and to provide absolute model-data fit evaluation, posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996) was used to assess the deviation of the model from the data. In PPMC, a test statistic is chosen to detect the systematic discrepancy between the observed data and the replicated data, and the posterior predictive *p*-value is computed through a comparison of the two test statistics over a large number of iterations. If an extreme *p*-value (close to 0 or 1) is observed, the fit of the model to the data is judged to be poor. In this study, two statistics were employed to assess the fit of the resulting model to the data. The proportion correct for each item was computed (i.e., classic item difficulty), and the mean of the proportion correct across items for each testing occasion was used to compute the posterior predictive *p*-value. Another statistic was used to compare the raw score distributions of the observed data and the predicted data. The examinees were divided into three groups (low, middle, and high scoring) on the basis of their raw scores (with cutoffs of 16 and 32 for a 48-item test, for example), and the number of examinees in each scoring group on each testing occasion was computed to evaluate the model fit using the posterior predictive *p*-value.

The PPMC results showed that the quadratic growth 3P-ML-HIRT model with Level-3 predictors yielded a good model fit because all the posterior predictive *p*-values were far from 0 or 1. Overall, it may be concluded that the proposed model fit the data well, that the tests covered difficult and easy items, that the three factor loadings were similar, and that a regional effect was observed.

Example 2: Assessment of Students' School Adaptability

In Taiwan, the number of female immigrants (mostly from Southeast Asia) has increased in recent decades because of marriage to local males. The school performance and adaptability of the offspring of these immigrants, who are often considered to have a higher probability of poor adaptability than other local students because of perceived cultural inferiority and biases, is a topic of interest. In the second polytomous item example, the adaptability of these immigrant students in elementary schools was investigated using the Students' School Adaptability inventory (Wu et al., 2008).

An inventory measuring Students' School Adaptability was administered to 565 elementary school students in Taiwan between 2005 and 2007. The inventory comprised three tests, assessing the teacher–student relationship (4 items), peer relationships (7 items), and the curriculum/environment (7 items). All of the tests were in the form of four-point polytomous items. All of the items were constant across the three testing times. The three tests were therefore treated as measuring first-order latent traits governed by a common second-order latent trait representing the overall adaptability.

Four questions related to ML-HIRT model selection were of specific interest. First, did the items share an identical slope parameter, and could a common set of threshold parameters be applied to all of the items (i.e., the RS-ML-HIRT model and the GR-ML-HIRT model)? Four common polytomous ML-HIRT models with linear growth were used to fit the data, and the resulting DIC values were 64,981.9 for the RS-ML-HIRT model, 64,513.2 for the GR-ML-HIRT model, 64,230.8 for the PC-ML-HIRT model, and 63,593.7 for the GPC-ML-HIRT model. The GPC-ML-HIRT model with linear growth therefore provided the best fit to the data. Second, the GPC-ML-HIRT model was compared to the GPC-ML-IRT model (i.e., all the test items measured a single latent trait) to determine whether the higher-order latent traits were necessary. The results showed that the GPC-ML-HIRT model yielded a better fit to the data (DIC = 63,593.7) than the corresponding ML-IRT model (DIC = 669,22.5) and that the three first-order latent traits can be governed by an overall performance of the school adaptability. Third, is a linear or nonlinear (quadratic) latent growth model more appropriate? Compared with the GPC-ML-HIRT model with quadratic growth (DIC = 63,596.1), the GPC-ML-HIRT model with linear growth fit the data better (DIC = 63,593.7). Finally, did the local and immigrant students display systematically different growth trajectories? Two linear growth GPC-ML-HIRT models with and without Level-3 predictors were compared, and the GPC-ML-HIRT model without Level-3 predictors (DIC = 63,593.7) provided the best fit, as its DIC value was smaller than that of the same model with Level-3 predictors (DIC = 63,596.4). In addition, the same PPMC methods used in the dichotomous-item example were adapted to assess the absolute model fit. The GPC-ML-HIRT model without Level-3 predictors was judged to provide a satisfactory fit to the data because all the posterior predictive *p*-values were substantially different from 0 or 1.

The estimates were between -2.82 and 1.86 (*M* = -0.87) for the location parameters and between 0.41 and 1.81 (*M* = 0.96) for the slope parameters. The estimates for the three factor loadings were 0.80, 1.11, and 1.00 for the teacher–student relationship, peer relationships, and the curriculum/environment, respectively. It may be concluded that a relatively wider range of location parameters was observed, that the three factor loadings were similar, and that the differences in the growth trajectories of the local and immigrant students were very small. In summary, the two examples illustrate successful applications of the ML-HIRT model to real data.