Article

Applied Psychological Measurement
2015, Vol. 39(5) 362–372
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614568112
apm.sagepub.com
SAGE

A Multilevel Higher Order Item Response Theory Model for Measuring Latent Growth in Longitudinal Data

Hung-Yu Huang[1]

Abstract

In educational and psychological testing, individuals are often repeatedly measured to assess the changes in their abilities over time or their latent trait growth. If a test consists of several subtests, the latent traits may have a higher order structure, and traditional item response theory (IRT) models for longitudinal data are no longer applicable. In this study, various multilevel higher order item response theory (ML-HIRT) models for simultaneously measuring growth in the second- and first-order latent traits of dichotomous and polytomous items are proposed. A series of simulations conducted using the WinBUGS software with Markov chain Monte Carlo (MCMC) methods reveal that the parameters could be recovered satisfactorily and that latent trait estimation was reliable across measurement times. The application of the ML-HIRT model to longitudinal data sets is illustrated with two empirical examples.

Keywords

item response theory, multilevel model, higher order model, latent growth, Bayesian estimation

In educational and psychological tests, individuals are often repeatedly measured on multiple occasions to determine how their abilities change over time (e.g., Andersen, 1985; Embretson, 1991). Item responses that are collected on multiple occasions comprise longitudinal data, and traditional estimation methods are not applicable to these data because the assumption of local item independence does not hold. In the factor analysis model literature, latent growth in examinee ability can be assessed using multilevel extensions for continuous and discrete outcomes (Muthén & Muthén, 2007). In these latent growth models, a single latent trait is measured using multiple indicators, and random effects are included to incorporate individual differences in both the initial status and the rate of growth. The diversity of item response functions imposes a serious constraint on the existing multilevel factor models (e.g., the Mplus computer program) because these models can only be applied to a limited number of specific functions. For example, the three-parameter logistic model (3PLM; Birnbaum, 1968) and other novel models (see De Boeck & Wilson, 2004) are not incorporated in the popular Mplus computer program. In item response theory (IRT), in contrast, a multilevel IRT model can accommodate a variety of

[1] University of Taipei, Taiwan

Corresponding Author:
Hung-Yu Huang, Department of Psychology and Counseling, University of Taipei, No. 1, Ai-Guo West Road, Taipei 10048, Taiwan.
Email: hyhuang@go.utaipei.edu.tw

item response functions for measuring latent trait changes in examinees (Hung & Wang, 2012). However, the existing multilevel IRT models have not been adapted to assessing latent trait growth in multiple latent traits.

A common assumption inherent to multilevel factor models and multilevel IRT models is that the latent trait intended to be measured is univariate. Multiple latent traits measured by multidimensional tests are very common in real testing situations, and measuring latent growth in multiple latent traits becomes increasingly interesting in longitudinal surveys. Several studies have proposed latent growth models for multidimensional tests (e.g., Bianconcini, 2012; Bollen & Curran, 2006; Raykov, 2007). However, these models require the consideration of a large number of random effects, use continuous indicators, and do not consider the higher order structure of latent traits.

Measuring changes within the framework of higher order latent traits is a complex task with multilevel IRT models because the extent to which each subtrait (i.e., first-order latent trait) contributes to the overall latent trait (i.e., second-order latent trait) is unknown. Accordingly, if simultaneous measurements of the latent growth in the first-order latent traits and the second-order latent trait are desired, then the weights of the second-order latent traits must be assigned to the first-order latent traits, and a higher order latent trait structure is necessary. In the context of IRT modeling, the incorporation of higher order latent traits into IRT models leads to higher order item response theory (HIRT) models that can accommodate a variety of item response functions (e.g., Huang, Wang, Chen, & Su, 2013). However, few of these HIRT models have been adapted to incorporate ability changes using multilevel extensions that fit longitudinal data. A new model that combines a higher order latent trait model with a multilevel model is therefore required to measure the growth in multiple latent traits with increased estimation efficiency.

In this study, a general class of multilevel higher order item response theory (ML-HIRT) models for longitudinal data is proposed in which the second- and first-order latent traits can be estimated simultaneously, the ability changes can be reliably assessed, the item response functions can be flexibly specified for dichotomous and polytomous items, and a linear or nonlinear latent growth model is plausible. This new class of ML-HIRT models not only integrates all of the existing HIRT models and multilevel IRT models but also provides new directions for future research.

This article is organized as follows. First, ML-HIRT models for dichotomous and polytomous items are developed and discussed. Second, simulations that were performed to assess the efficiency of parameter estimation in the ML-HIRT models are presented. Third, the application of ML-HIRT models to longitudinal surveys is demonstrated through two empirical examples involving dichotomous and polytomous items. Finally, conclusions are drawn concerning the new models, and suggestions for future research are provided.

The ML-HIRT Model

A ML-HIRT model consists of (but is not limited to) three levels, with each level constituting a specific model: a within-occasion model (Level 1), a within-person model (Level 2), and a between-person model (Level 3). The within-occasion model describes item responses for specific occasions, the within-person model addresses variations in latent traits over a single person's measurement occasions, and the between-person model specifies the variations in growth trajectories between persons.

For simplicity, a two-order structure with one common second-order latent trait is assumed. In this structure, each item is governed by a first-order latent trait. At Level 1, a specified item response model can be formulated. For example, the three-parameter multilevel higher order IRT (3P-ML-HIRT) model can be expressed as follows:

$$P_{nti1v} = \pi_{iv} + (1-\pi_{iv})\times \frac{\exp\left[\alpha_{iv}\left(\theta^{(1)}_{ntv}-\delta_{iv}\right)\right]}{1+\exp\left[\alpha_{iv}\left(\theta^{(1)}_{ntv}-\delta_{iv}\right)\right]}, \tag{1}$$

where

$$\theta^{(1)}_{ntv} = \lambda_v \theta^{(2)}_{nt} + \varepsilon^{(1)}_{ntv}; \tag{2}$$

α_{iv} is the slope (discrimination) parameter; δ_{iv} is the location (difficulty) parameter; π_{iv} is the asymptotic (pseudo-guessing) parameter for item i in test v; $\theta^{(2)}_{nt}$ and $\theta^{(1)}_{ntv}$ are the second-order and the vth first-order latent traits, respectively, for person n at time t; λ_v is the regression weight (factor loading) specifying the relationship between the second-order latent trait and the vth first-order latent trait; and $\varepsilon^{(1)}_{ntv}$ is the residual of time t in test v for person n and is assumed to be normally distributed with a mean of zero and to be independent of the other εs and θs. Note that the factor loadings and the item parameters do not have the subscript t, indicating that these parameters do not change over time. For model identification, one of the regression weights is set to a fixed value (of one), and one of the discrimination parameters in test v is set to one to allow for different tests using a common metric.

The within-person model can be formulated at Level 2, and a latent growth model is defined by

$$\theta^{(2)}_{nt} = \boldsymbol{\omega}_{nt}\boldsymbol{\beta}_n + \psi_{nt}, \tag{3}$$

with

$$\psi_{nt} \sim N\left(0, \sigma^2_{\psi(t)}\right), \tag{4}$$

where $\boldsymbol{\beta}'_n = [\beta_{n0}, \beta_{n1}, \ldots, \beta_{nh}]$ is a vector of length $h+1$ for the growth factors for person n that specify initial statuses and growth rates; $\boldsymbol{\omega}_{nt} = [1, \omega^1_{nt}, \omega^2_{nt}, \ldots, \omega^h_{nt}]$ is a vector of time-based loadings; h can be specified for the polynomial growth curve ($h = 1$ for a linear latent growth model); and ψ_{nt} is the regression residual for person n at time t, which is assumed to be independent of the other εs and ψs.

Finally, to specify the variation in the growth trajectories of individuals, the coefficients in Equation 3 can be regressed against another set of personal background variables (the number of predictors is set to m) to formulate the Level 3 model, which is defined by

$$\beta_{nd} = \boldsymbol{\kappa}_n \boldsymbol{\gamma}_d + \varsigma_{nd}, \tag{5}$$

with

$$\varsigma_n \equiv [\varsigma_{n0}, \varsigma_{n1}, \ldots, \varsigma_{nh}] \sim N(0, \boldsymbol{\Sigma}_\varsigma), \tag{6}$$

where $\boldsymbol{\kappa}_n$ is a set of observed predictors for person n; $\boldsymbol{\gamma}'_d = [\gamma_{d0}, \gamma_{d1}, \ldots, \gamma_{dm}]$ is the regression coefficient parameter vector (including the intercept term) for the dth coefficient at Level 2; $d = 0, 1, \ldots, h$; $\boldsymbol{\varsigma}_n$ is the vector of Level 3 regression residuals (which are assumed to be mutually independent of the Level 2 and Level 1 residuals); and $\boldsymbol{\Sigma}_\varsigma$ is a variance–covariance matrix of dimension $h+1$. For model identification, the intercept term in vector $\boldsymbol{\gamma}_0$, with β_{n0} as the criterion variable, is set to zero. When more than three levels are involved, the proposed model can be readily extended; an example of a four-level two-order IRT model is provided in the Online Appendix A. To simplify the situation, this study focused only on the three-level HIRT model and the linear growth model.

Other dichotomous IRT models can be extended to a multilevel approach in a straightforward manner. If a two-parameter logistic model (2PLM) or a one-parameter logistic model (1PLM) with higher order latent traits is used as the item response function at Level 1, then a two-parameter multilevel higher order IRT (2P-ML-HIRT) model and a one-parameter multilevel higher order IRT (1P-ML-HIRT) model can be formulated.

Huang et al. (2013) developed the polytomous HIRT model and proposed four commonly used item response models for use with polytomous items in the presence of higher order latent traits. In the framework of ML-HIRT models, for a polytomous item measuring a first-order latent trait (as in the generalized partial credit model [GPCM], Muraki, 1992, for example) at Level 1, the log odds can be defined by

$$\log\left(\frac{P_{ntijv}}{P_{nti(j-1)v}}\right) = \alpha_{iv}\left(\theta_{ntv}^{(1)} - \eta_{iv} - \tau_{ijv}\right), \tag{7}$$

where P_{ntijv} and $P_{nti(j-1)v}$ are the probabilities of scoring j and $j-1$ on item i in test v at time t for person n, η_{iv} is the overall difficulty of item i in test v, τ_{ijv} is the jth threshold parameter of item i in test v, and the other parameters are as defined above. Combining Equations 2 and 7 leads to the GPC-ML-HIRT. If the overall difficulty parameter and the threshold parameter for item i and threshold j in test v are combined, the δ_{ijv} location parameter arises. Similarly, if the item response function follows the partial credit model (PCM), the rating scale model (RSM), and the graded response model (GRM), then the corresponding ML-HIRT models are denoted by PC-ML-HIRT, RS-ML-HIRT, and GR-ML-HIRT, respectively.

In the formulation of the ML-HIRT model described above, it is assumed that a test is repeatedly administered to individuals and that all of the item parameter values are identical over time (i.e., these item parameters do not have t subscripts). Over time, however, it is possible for a new set of items that were not present in the earlier tests to include parts of items that earlier tests may have excluded. In such cases, a test measuring the same latent trait consists of different items on different occasions. As long as a sufficient common set of items are used as an anchor over time (four common items between any two occasions; see Wang, 2004), the ML-HIRT model can calibrate all of the item parameters with respect to the common metric, and the notation of the ML-HIRT item parameters can be incorporated into the subscript t to indicate that the items were not administered on all of the occasions.

Method

Simulation Design

Two simulations were performed in this study: one for dichotomous items and the other for polytomous items. In both cases, the simulation design and generated values were chosen to be consistent with the two empirical studies (Chang, 2007; Wu, Tsai, & Siao, 2010), and the details of the design of the empirical study are described in the Online Appendix F. Three first-order latent traits (i.e., three tests) and one second-order latent trait (overall ability) were assumed. Note that different measurement occasions may have different items and that a sufficient set of anchor items were used to build the common metric.

The 3P-ML-HIRT model with linear growth was used to generate simulated data, with each data set consisting of 4,000 individuals (2,400 individuals residing in an urban region and 1,600 individuals residing in a rural region). A smaller data set of 1,000 (600 urban examinees and

400 rural examinees) was also simulated to assess the effect of sample size on the parameter estimation. The regions of students were used as the predictors at Level 3. A MATLAB computer program was written by the author to generate item responses. The true values for the model parameters are summarized in Table 1, and the simulation process is documented in the Online Appendix B. Thirty replications were performed to assess the dichotomous parameter recovery in the ML-HIRT model. Note that all of the items were assumed to share a common pseudo-guessing parameter because of the uncertainty in this parameter.

In the second simulation, the GPC-ML-HIRT model with linear growth was used to generate item responses for the polytomous items. The simulation design was based on the following empirical polytomous item example using a linear growth model with Level 3 predictors, as shown in the Online Appendix F. Three tests measuring a first-order latent trait were generated, and a common second-order latent trait was assumed to govern these three first-order latent traits. The values generated are summarized by their ranges or are listed as their true values in Table 4. Thirty replications, each with 565 simulated individuals, were conducted. In addition, two groups of local children (278 students, coded as 1) and immigrant children (287 students, coded as -1) were used as the predictors at Level 3 to determine the difference between the growth trajectories of these two groups.

Analysis

Because of the high dimensionality of the model, the WinBUGS freeware program (Spiegelhalter, Thomas, & Best, 2003) with Markov chain Monte Carlo (MCMC) methods was used to calibrate the model parameters. The prior settings and parameter convergence evaluation are present in the Online Appendix C. The WinBUGS commands for the proposed models are available upon request. The Bayesian deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) was used to assess the fit of the model to the data. A smaller DIC value indicates a better fit of the model to the data.

For each estimator, the bias and root mean square error (RMSE) were computed to assess the parameter recovery in the simulations. The absolute value of relative bias (ARB) was also computed. If ARB is close to zero, the estimator is considered acceptable (Hoogland & Boomsma, 1998). In addition, the square root of the ratio of the mean error variance to the sampling variance was computed to assess the accuracy of the estimated standard errors for each parameter estimate. If the estimated standard error was close to the empirical standard error, then the ratio would approach unity, and the estimated standard error was considered appropriate for statistical hypothesis testing. It was expected that the model parameters could be recovered well for the proposed models, that the estimated standard errors would be appropriate, and that these models could be successfully applied to measuring latent growth or ability changes in real data.

Results

Parameter Recovery for Dichotomous Items

Table 1 summarizes the parameter recovery for the dichotomous items and the 3P-ML-HIRT model for the sample size of 4,000, which corresponds to the within-occasion model (Level 1) for the 3P-HIRT item response function. The bias magnitudes were quite small, but the factor loadings were slightly overestimated. Most of the mean RMSE values for the different types of parameters were less than 0.1, and the parameter recovery was acceptable. Most estimators had ARB values between 0 and 0.1, suggesting that the parameter recovery was marginally satisfactory. Several estimators had ARB values greater than 0.1 because of the parameters' near-zero

Table 1. Parameter Recovery for the 3P-ML-HIRT Model With a Sample Size of 4,000.

Parameter	True	Bias	RMSE	ARB	Ratio	Skewness
Difficulty						
M	0.257	0.016	0.117	0.095	0.988	0.006
SD	1.297	0.089	0.082	0.210	0.163	0.427
Maximum	3.367	0.261	0.461	2.550	1.386	1.043
Minimum	−4.224	−0.229	0.029	0.001	0.655	−1.320
Slope						
M	1.001	−0.053	0.090	0.054	0.977	0.058
SD	0.428	0.042	0.048	0.031	0.139	0.418
Maximum	2.371	0.039	0.275	0.108	1.339	1.447
Minimum	0.244	−0.154	0.019	0.000	0.730	−0.628
Pseudo-guessing	0.147	0.001	0.003	0.003	1.242	−0.241
Residual variance at Level 1						
M	0.246	0.043	0.066	0.191	0.936	0.280
SD	0.263	0.049	0.056	0.175	0.178	0.436
Maximum	1.069	0.181	0.216	0.693	1.228	0.950
Minimum	0.015	−0.003	0.016	0.012	0.681	−0.646
Residual variance at Level 2						
ψ_1	0.028	0.001	0.009	0.025	0.853	0.163
ψ_2	0.036	0.002	0.006	0.042	1.148	−0.314
ψ_3	0.046	0.000	0.012	0.002	0.741	0.231
ψ_4	0.021	0.007	0.015	0.324	0.759	1.019
Residual variance at Level 3						
Intercept (β_0)	0.428	0.003	0.034	0.006	0.812	0.174
Slope (β_1)	0.039	0.002	0.005	0.046	0.874	0.301
Covariance	0.073	0.003	0.009	0.045	0.756	0.501
Regression weight at Level 3						
Intercept of β_0	0	NA	NA	NA	NA	NA
Slope of β_0	0.190	0.003	0.012	0.015	1.116	−0.151
Intercept of β_1	0.396	0.006	0.017	0.015	0.794	−0.356
Slope of β_1	0.032	0.002	0.005	0.059	1.116	1.146
Factor loading						
λ_1	1	NA	NA	NA	NA	NA
λ_2	1.129	0.124	0.181	0.110	0.663	0.613
λ_3	1.494	0.104	0.135	0.069	0.964	0.856

Note. The number of estimated difficulty parameters and estimated slope parameters and the estimated Level 1 residual variances are 105, 102, and 12, respectively. 3P-ML-HIRT = three-parameter multilevel higher order item response theory; True = true values; RMSE = root mean square error; ARB = absolute value of relative bias; NA = not applicable because of model constraints.

true values (e.g., $\delta = 0.002$ corresponded to the greatest ARB value of 2.550). Parameter values were estimated with considerably less accuracy for certain items included in the set of items that were unique over time. The parameter recoveries for different times and for the unique and common items are shown in the Online Appendix D. In addition, the estimated standard errors were appropriate because most of the square roots of the ratios of estimated error variances to sampling variances were close to unity. Most of the estimates yielded symmetric distributions, with computed skewness values close to zero.

The quality of the parameter recovery was slightly lower when the sample size was decreased to 1,000. As shown in Table 2, a smaller sample size led to larger bias and RMSE values. Although the larger ARB statistics were associated with near-zero true values, all the estimators had larger ARB values for the sample size of 1,000 than for the sample size of 4,000. The

Table 2. Parameter Recovery for the 3P-ML-HIRT Model With a Sample Size of 1,000.

Parameter	True	Bias	RMSE	ARB	Ratio	Skewness
Difficulty						
M	0.257	0.031	0.228	0.502	1.215	−0.114
SD	1.297	0.210	0.156	3.536	0.314	0.451
Maximum	3.367	0.593	0.701	44.70	2.393	1.046
Minimum	−4.224	−0.569	0.056	0.001	0.678	−1.223
Slope						
M	1.001	−0.111	0.152	0.115	1.278	0.121
SD	0.428	0.087	0.081	0.062	0.355	0.581
Maximum	2.371	0.111	0.370	0.246	3.419	1.915
Minimum	0.244	−0.344	0.031	0.000	0.642	−1.219
Pseudo-guessing	0.147	0.004	0.008	0.024	0.991	−0.935
Residual variance at Level 1						
M	0.246	0.110	0.149	0.512	1.105	0.118
SD	0.263	0.122	0.126	0.565	0.337	0.465
Maximum	1.069	0.438	0.480	2.207	1.625	0.985
Minimum	0.015	0.003	0.026	0.034	0.577	−0.817
Residual variance at Level 2						
ψ_1	0.028	0.008	0.021	0.039	2.028	−0.728
ψ_2	0.036	0.007	0.011	0.197	0.889	0.580
ψ_3	0.046	0.018	0.036	0.185	0.788	0.407
ψ_4	0.021	−0.001	0.007	0.852	0.674	1.842
Residual variance at Level 3						
Intercept (β_0)	0.428	0.028	0.056	0.066	1.212	−0.538
Slope (β_1)	0.039	0.004	0.011	0.113	1.381	−0.589
Covariance	0.073	0.004	0.007	0.049	1.173	−0.523
Regression weight at Level 3						
Intercept of β_0	0	NA	NA	NA	NA	NA
Slope of β_0	0.190	0.003	0.019	0.015	1.399	0.326
Intercept of β_1	0.396	0.017	0.026	0.043	1.225	−0.033
Slope of β_1	0.032	0.004	0.011	0.131	1.001	−0.488
Factor loading						
λ_1	1	NA	NA	NA	NA	NA
λ_2	1.129	0.287	0.309	0.254	1.431	0.298
λ_3	1.494	0.171	0.182	0.114	2.580	0.595

Note. The number of estimated difficulty parameters and estimated slope parameters and the estimated Level 1 residual variances are 105, 102, and 12, respectively. 3P-ML-HIRT = three-parameter multilevel higher order item response theory; True = true values; RMSE = root mean square error; ARB = absolute value of relative bias; NA = not applicable because of model constraints.

estimated standard errors for the smaller sample size were slightly less accurate than those for the larger sample size. The skewness patterns of the estimates across replications differed little between the two sample sizes; most of the skewness values were between −1 and 1, suggesting that the estimates followed a symmetric distribution. In summary, the sample size was associated with both the quality of the parameter recovery and the accuracy of standard error estimates but had little impact on the skewness of the estimates.

Table 3 lists the mean RMSE values for the person parameter estimates for the four testing occasions across replications for the second-order latent trait and for the three first-order latent traits. As Table 3 shows, for the larger sample size of 4,000, the second-order latent trait could be recovered better than the three first-order latent traits, and person parameter estimation became poor as time progressed because a large number of individuals with extremely atypical

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

Condition	Occasion 1	Occasion 2	Occasion 3	Occasion 4
Sample Size of 4,000				
First-order latent trait in Test 1	0.269	0.354	0.431	0.438
First-order latent trait in Test 2	0.510	0.498	0.584	0.632
First-order latent trait in Test 3	0.416	0.423	0.571	0.762
Second-order latent trait	0.245	0.246	0.300	0.362
Sample size of 1,000				
First-order latent trait in Test 1	0.274	0.359	0.437	0.450
First-order latent trait in Test 2	0.556	0.575	0.701	0.810
First-order latent trait in Test 3	0.436	0.460	0.623	0.830
Second-order latent trait	0.248	0.252	0.309	0.374

Note. RMSE = root mean square error; 3P-ML-HIRT = three-parameter multilevel higher order item response theory.

abilities were observed on later occasions. Patterns similar to those for the large sample size were observed for the small sample size of 1,000, although the small sample size resulted in slightly inferior person parameter recovery, as shown in the bottom of Table 3. The poor estimation for the later occasions may have occurred because the ranges of the item difficulty parameters did not correspond exactly to the levels of the examinees' abilities; the extreme ends of the ability range could not be recovered well. An additional test of the hypothesis that the ranges of the item parameter values affect the estimation precision for the examinees is shown via simulation in the Online Appendix E. In summary, the 3P-ML-HIRT model yielded marginally acceptable estimates of both the second-order and first-order latent traits, and the estimation was improved when a wider range of item difficulty parameters was adopted. This improvement can be achieved through the involvement of testing and content experts in the further development of the tests.

Parameter Recovery for Polytomous Items

The parameter recovery results for the polytomous items using the GPC-ML-HIRT model, corresponding to the within-occasion model of the GPC-HIRT item response function, are presented in Table 4. All of the bias values were so close to zero and the RMSE values were so small that the GPC-ML-HIRT model yielded good parameter estimation. Most of the estimators had marginally acceptable ARB statistics, with values below 0.1. The largest ARB value for item difficulty (1.495) corresponded to a near-zero true value (0.019). The estimated standard errors were judged appropriate because most of the ratios were very close to unity. The assessment of the skewness of the parameter estimates across replications showed that the estimates yielded a symmetric pattern and that the distribution was not severely skewed. In summary, the parameter recovery for the polytomous items using the GPC-ML-HIRT model appeared to be good, and the standard errors obtained by MCMC estimation were appropriate.

The assessment of the latent trait parameter recovery for the GPC-ML-HIRT model for the three occasions showed that the mean RMSE values were 0.529, 0.496, and 0.527, respectively, for Test 1; 0.555, 0.515, and 0.564, respectively, for Test 2; 0.512, 0.504, and 0.501, respectively, for Test 3; and 0.472, 0.434, and 0.467, respectively, for the second-order latent trait. The estimation for the second-order latent trait outperformed the estimation for the three first-order latent traits in terms of the recovery of person parameters. The principal reason for the slightly higher RMSE values for both orders was that the number of test items was not

Table 4. Parameter Recovery for the GPC-ML-HIRT Model With a Sample Size of 565.

Parameter	True	Bias	RMSE	ARB	Ratio	Skewness
Location						
M	−0.875	0.003	0.125	0.068	1.131	−0.241
SD	1.114	0.040	0.044	0.200	0.144	0.347
Maximum	1.846	0.093	0.246	1.495	1.639	0.675
Minimum	−2.834	−0.087	0.051	0.000	0.924	−1.066
Slope						
M	0.970	−0.031	0.083	0.031	1.054	0.341
SD	0.330	0.025	0.032	0.016	0.159	0.422
Maximum	1.816	−0.001	0.173	0.066	1.477	1.026
Minimum	0.412	−0.101	0.033	0.001	0.788	−0.273
Residual variance at Level 1						
M	0.662	0.051	0.130	0.082	1.069	0.378
SD	0.253	0.022	0.041	0.038	0.161	0.340
Maximum	1.310	0.087	0.226	0.151	1.328	0.907
Minimum	0.417	0.024	0.072	0.035	0.850	−0.117
Residual variance at Level 2						
ψ_1	0.361	0.030	0.118	0.082	0.880	−0.588
ψ_2	0.241	0.030	0.062	0.124	1.203	−0.053
ψ_3	0.093	0.023	0.064	0.242	1.169	1.023
Residual variance at Level 3						
Intercept (β_0)	0.407	0.052	0.119	0.127	0.868	0.278
Slope (β_1)	0.060	0.006	0.026	0.107	1.204	1.009
Covariance	0.014	0.001	0.040	0.071	0.992	−0.724
Regression weight at Level 3						
Intercept of β_0	0	NA	NA	NA	NA	NA
Slope of β_0	0.112	0.013	0.051	0.113	0.886	−0.164
Intercept of β_1	−0.148	0.008	0.026	0.052	1.081	0.001
Slope of β_1	0.010	−0.006	0.030	0.560	0.889	−0.229
Factor loading						
λ_1	0.807	−0.011	0.072	0.014	1.086	0.409
λ_2	1.123	−0.037	0.103	0.033	1.006	0.305
λ_3	1	NA	NA	NA	NA	NA

Note. The number of estimated location parameters and estimated slope parameters and the estimated Level 1 residual variances are 18, 15, and 9, respectively. GPC-ML-HIRT = generalized partial credit–multilevel higher order item response theory; True = true values; RMSE = root mean square error; ARB = absolute value of relative bias; NA = not applicable because of model constraints.

sufficiently large to provide precision estimates for person parameters. Compared with the results for dichotomous items, the GPC-ML-HIRT model yielded relatively consistent estimates for the three occasions; the range of ability levels was more consistent with that of the item location parameters for the polytomous items than with the range for the dichotomous items. In summary, the GPC-ML-HIRT model yielded marginally acceptable person parameter estimates for both orders over time. The applications of the ML-HIRT model to empirical data with dichotomous and polytomous items are shown in the Online Appendix F.

Discussion and Conclusion

In this study, a multilevel extension of the HIRT model (the ML-HIRT model) was developed for longitudinal data, and the efficiency of the new model was evaluated through simulation studies and empirical analyses. The results of a series of simulations using WinBUGS with MCMC methods indicated that the model parameters can be recovered fairly well, that the standard errors were estimated appropriately, and that both orders of individual latent traits can be estimated satisfactorily over time. As the sample size increases, the item parameter estimation becomes more precise. The use of higher and lower difficulty parameters improved the precision of the estimation of the person parameters for the later measurement occasions. In practice, apart from the simulation study, it is difficult to know a priori how wide a range of item difficulty parameters the tests require unless a pilot survey is conducted; sequential testing and input from content experts can be used to adjust the setting. In addition, prior information from other surveys (e.g., the large-scale assessment of the Programme for International Student Assessment [PISA]) can assist psychometricians and testing writers to design substantially more difficult or easier items to yield a wider range of item difficulty parameters. Note that the performance of the estimators was poor (to a slight extent) for certain items and that improved parameter estimates can be obtained by including more common items to serve as an anchor over time (discussed below) and by increasing the sample size, as demonstrated in the simulation. Although the simulations were designed to resemble empirical settings, the manipulation of other factors, such as the test length and the number of subtests, deserves further investigation.

It is not uncommon in longitudinal surveys for different sets of items to be used to measure the same latent trait in tests administered on different occasions. In this case, identifying the anchor items is a prerequisite for linking different metrics over time to estimate model parameters precisely. In the dichotomous-item simulation described in this article, common items in each test were used as anchors across time, and most of the parameter recovery was acceptable, although the factor loadings were slightly overestimated. In an additional simulation whose topic was the use of the tests for Occasion 1 for the other three occasions (i.e., simulating the administration of the same tests on all four occasions) for a sample size of 4,000, the parameter estimation was improved. More common items are therefore recommended; however, the effect of practice when the same test is administered multiple times is inevitable, especially in achievement assessment.

The ML-HIRT model is quite flexible and can easily accommodate more than three levels or a polynomial growth curve model. As described in this article, a between-class model (Level 4) and a between-school model (Level 5) can be embedded in the multilevel structure, and the use of a polynomial growth curve model in place of a linear latent growth model can incorporate the latent growth trajectories. As for the higher order latent trait model, more than two orders of latent traits or two (or more) second-order latent traits can be introduced, and the relationship between the second- and first-order latent traits can be nonlinear (e.g., quadratic). The analogous extension of the ML-HIRT model is left for future research.

For decades, the detection of differential item functioning (DIF) or measurement invariance over groups has been the focus of the development of new tests. In the ML-HIRT model, item parameters and factor loadings may be different between groups, and these parameters may shift as time progresses. An ML-HIRT model with fewer constraints may provide a promising approach to the detection of possible violations of measurement invariance or the assessment of DIF in the development of tests for longitudinal surveys. How to extend the ML-HIRT model to address this concern requires further investigation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partly supported by the Ministry of Science and Technology (102-2410-H-845-003-MY2).

Supplemental Material

The online appendixes are available at http://apm.sagepub.com/supplemental

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3-16.

Bianconcini, S. (2012). A general multivariate latent growth model with applications to student achievement. *Journal of Educational and Behavioral Statistics, 37*, 339-364.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees' ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York, NY: John Wiley.

Chang, L.-Y. (2007). *Taiwan education panel survey: Users' guide and base year (2007) student questionnaire for junior high school*. Taipei, Taiwan: Center for Survey Research, Academia Sinica.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-515.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*, 329-367.

Huang, H.-Y., Wang, W.-C., Chen, P.-H., & Su, C.-M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement, 37*, 619-637.

Hung, L.-F., & Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics, 37*, 231-255.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, L., & Muthén, B. (2007). *Mplus user's guide* (4th ed.). Los Angeles, CA: Author.

Raykov, T. (2007). Longitudinal analysis with regressions among random effects: A latent variable modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 146-169.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583-616.

Spiegelhalter, D. J., Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.

Wu, Y.-Y., Tsai, C.-C., & Siao, R.-F. (2010). Parental involvement of Southeastern Asian female immigrants and its relationship to their children's school life adjustments. *Journal of Research in Education Sciences, 55*, 157-186.