Article

A Multilevel Longitudinal Nested Logit Model for Measuring Changes in Correct Response and Error Types

Applied Psychological Measurement
2018, Vol. 42(1) 73–88

DOI: 10.1177/0146621617703182
journals.sagepub.com/home/apm
SAGE

Youngsuk Suh[1], Sun-Joo Cho[2], and Brian A. Bottge[3]

Abstract

This article presents a multilevel longitudinal nested logit model for analyzing correct response and error types in multilevel longitudinal intervention data collected under a pretest–posttest, cluster randomized trial design. The use of the model is illustrated with a real data analysis, including a model comparison study regarding model complexity and cluster bias. Two substantive research questions regarding the intervention effect on correct response probability and error patterns are investigated using the proposed model. The recovery of item parameters for the proposed model using two sample size conditions is examined via a simulation study. The accuracy of the parameter estimates is comparable with those found in previous studies for the same family of models, except for the intercept parameters of correct responses. Finally, the impact of ignoring cluster membership in the model on the parameter estimation is also studied by fitting a single-level model to multilevel data. Ignoring cluster membership in the model adversely affects the estimation of intercept parameters in correct and error responses.

Keywords

multilevel longitudinal model, nested logit model, Bayesian data analysis, error analysis.

Cluster randomized trials are experiments in which clusters of individuals, rather than independent individuals, are randomly allocated to intervention groups (Raudenbush, 1997). Due to the nature of the randomized clusters, this design forms a multilevel structure and thus possibly introduces dependence among individuals sampled within the same cluster. Therefore, when analyzing differences in outcomes between treatment and control groups, researchers typically use hierarchical linear modeling (HLM) to model this dependence.

Among others, a pretest–posttest, cluster randomized trial design is one of the most common designs used in practice, and this design produces multilevel longitudinal data. Reflecting the

[1]Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
[2]Vanderbilt University, Nashville, TN, USA
[3]University of Kentucky, Lexington, KY, USA

Corresponding Author:
Youngsuk Suh, Department of Educational Psychology, Graduate School of Education, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901, USA.
Email: yssuh327@gmail.com

popularity of item response theory (IRT) model applications in educational and psychological testing programs, several researchers have extended IRT models for dealing with multilevel structures, called multilevel IRT models (e.g., Fox & Glas, 2001; Kamata, 2001; Mislevy & Bock, 1989; Rabe-Hesketh, Skrondal, & Pickles, 2004; Raudenbush & Bryk, 2002), and a marginal maximum likelihood (MML) estimation and a Bayesian estimation have been used to estimate model parameters (see Fox & Glas, 2016, for a detailed review of multilevel IRT models and parameter estimation).[1] The multilevel IRT models enable researchers to study the impact of different predictors and response patterns in each level of multilevel data (e.g., Fox & Glas, 2001). When the multilevel structure is ignored, aggregation bias (also known as the ecological fallacy) may occur, and bias in estimated measurement precision would be expected (e.g., Raudenbush & Bryk, 2002).

In addition, IRT models have been applied to longitudinal categorical data (e.g., Andersen, 1985; Cai, 2010; Embretson, 1991). The main purposes of these models are to model latent variable(s) at each time point or to model changes between time points. Recently, multilevel extensions of longitudinal IRT models—specifically, multilevel longitudinal IRT models—were presented by B. Muthén and Asparouhov (2016) and von Davier, Xu, and Carstensen (2011). These multilevel longitudinal IRT models are intended to accurately measure individual differences at each time point (B. Muthén & Asparouhov, 2016) or individual differences in changes over time and to detect group differences (von Davier et al., 2011), by accounting for the multilevel data structure. Acquiring more accurate, individual differences in changes over time ensures better assessment of the intervention effect, which is often the major interest in a cluster randomized trial design.

When categorical outcomes, such as item responses for multiple-choice items and constructed-response items, are collected to measure an intervention effect, either dichotomous or polytomous IRT models can be fitted to the data, depending on the scoring scheme of the responses. Although binary scoring—that is, modeling correct responses using dichotomous IRT models—is a more common practice, modeling incorrect responses (e.g., distractors in multiple-choice items and error types in constructed items) in addition to the correct responses can be desirable to improve measurement accuracy. Research on IRT has highlighted the potential value of modeling polytomous responses, as opposed to binary responses, for purposes such as estimating the ability parameter (e.g., Baker & Kim, 2004), recovering linking coefficients (e.g., Kim, 2006), and examining differential distractor functioning (e.g., Suh & Bolt, 2011).

Analyzing incorrect responses (e.g., error types) has become an active research area, particularly in mathematics education (e.g., Charalambous & Pitta-Pantazi, 2007; Clarke & Roche, 2009), as this analysis can (a) identify the patterns of errors or mistakes that students make in their work, (b) explain why students make the errors, and (c) provide targeted instructions to correct these errors. Especially when clusters are the unit of randomization, as typically is the case in data sets from cluster randomized trials, treatment assessment can be expensive. Therefore, it is critical to extract additional information from item response data to improve assessment efficiency.

Despite the practical importance of using the information from incorrect responses, the existing extensions of multilevel longitudinal IRT models (e.g., B. Muthén & Asparouhov, 2016) are limited to modeling correct responses. In addition, current practices for analyzing incorrect responses or error types in multilevel longitudinal data rely on non-IRT-based analyses, using, for instance, a frequency table for error types or adopting HLM for error analysis (e.g., Bottge, Ma, Gassaway, Butler, & Toland, 2014). HLM analyses are often conducted with sum scores (or total scores). Based on previous research, a major concern of using total scores in HLM is the measurement error in the sum scores. Measurement error in the sum scores is responsible for biased parameter estimates and loss of power for detecting relationships among variables

(e.g., Fox & Glas, 2003). One method for alleviating this concern is using latent variable models (e.g., Cole & Preacher, 2014), such as IRT models.

Before analyzing multilevel longitudinal data to make inferences about the intervention effect, whether the construct of interest is the same across all levels (i.e., individual and cluster levels) must be checked. This can be examined by testing whether the item discriminations are equal across all levels. If the item discriminations are equal, then the model formulation with the same discrimination across all levels assumes one factor, one latent variable. In this case, the latent variable at the cluster level can be interpreted as the cluster mean of the latent variable at the individual level. If the item discriminations are not equal, then the model with different item discriminations over the respective individual and cluster levels cannot assume a common latent variable, and consequently, the individual- and cluster-level latent variables cannot be interpreted in the same way across all levels (B. Muthén, 1990). If this occurs, it is often referred to as cluster bias (Jak, Oort, & Dolan, 2013). For example, when test data come from students nested within teachers to measure teacher effectiveness, cluster bias means that the factors may be different over levels, such as students' ability at the student level and teaching strategy at the teacher level (as the characteristics of classrooms). A simulation study showed that the item discrimination estimates and standard errors of a (cross-sectional) multilevel IRT model were biased when cluster bias was ignored (Lee & Cho, 2015). The biased item discrimination estimates can result in the misinterpretation of latent variables and biased IRT scores.

The purpose of the present study is twofold: to specify an IRT model for analyzing the correct response and error types in multilevel longitudinal intervention data collected under a pretest–posttest, cluster randomized trial design and to illustrate the use of the specified model through a real data analysis. In the real data analysis, a series of analyses are conducted using the model. First, a model comparison analysis for choosing the best-fitting model is described among candidate models: that is, to determine whether the multilevel longitudinal model is preferred over the longitudinal model and whether a measurement model with cluster bias is preferred over a model without cluster bias. Second, measurement invariance across groups and time points is briefly examined as a preliminary analysis before inferences are made about the intervention effects. Then, the authors illustrate how the model can be used to detect the intervention effects by examining the changes in the probabilities of correct response and each error type after the intervention. Two substantive research questions are investigated in an instructional intervention study using the specified model. Finally, the recovery of item parameters for the proposed model and the impact of ignoring cluster membership in the model on the parameter recovery are examined via a simulation study.

Multilevel Longitudinal Nested Logit Model

When item responses are nominally scored, perhaps the most common IRT approach is reflected by Bock's (1972) nominal response model (NRM), which is a divide-by-total model (Thissen & Steinberg, 1986). As an alternative to the NRM, Suh and Bolt (2010) proposed two-parameter logistic (2PL) nested logit model (NLM) and three-parameter logistic (3PL) NLM, referred to as 2PL-NLM and 3PL-NLM, respectively. Unlike existing models (the divide-by-total models), the models possess the attractive statistical property of *category collapsibility* regarding incorrect responses. This property allows direct evaluation of the probabilities of incorrect responses, independent of the correct response probability. That is, the NLMs use a traditional functional form for the correct response, such as 2PL probability, while investigating the functions of incorrect responses in contexts in which

practitioners might ultimately intend to use binary item scoring. This property also provides advantages for examining measurement invariance issues in correct responses and incorrect responses separately (e.g., Suh & Bolt, 2011). In this study, the 2PL-NLM is adopted as a measurement model for analyzing multilevel longitudinal data.

Suppose a test is composed of I items and each item has one correct response category and M incorrect response categories. Let y_{ij} represent an item response for item i by examinee j once keyed for correctness (i.e., y_{ij} = 1 if correct, and 0 otherwise). Furthermore, let d_{ijv} denote an item response for incorrect response category v ($v = 1, 2, \ldots, M$) such that $d_{ijv} = 1$ when examinee j shows incorrect category v of item i, and 0 otherwise. Under the 2PL-NLM (Suh & Bolt, 2010), the probability that an examinee who has ability θ_j chooses the correct response category for item i is modeled as a traditional 2PL model:

$$P(y_{ij} = 1|\theta_j) = \frac{\exp(\beta_i + \alpha_i\theta_j)}{1 + \exp(\beta_i + \alpha_i\theta_j)}, \tag{1}$$

where β_i denotes the intercept parameter, and α_i is the slope parameter for item i. The probability that the examinee shows incorrect response category v is modeled as the product of the probability of getting the item wrong and the probability of selecting incorrect category v:

$$\begin{aligned} P(y_{ij} = 0, d_{ijv} = 1|\theta_j) &= P(y_{ij} = 0|\theta_j)P(d_{ijv} = 1|y_{ij} = 0, \theta_j) \\ &= \left[1 - \frac{\exp(\beta_i + \alpha_i\theta_j)}{1 + \exp(\beta_i + \alpha_i\theta_j)}\right] \left[\frac{\exp(\zeta_{iv} + \lambda_{iv}\theta_j)}{\sum_{k=1}^{M} \exp(\zeta_{ik} + \lambda_{ik}\theta_j)}\right], \end{aligned} \tag{2}$$

where ζs and λs are the intercept and slope parameters for the incorrect responses, respectively. The second bracketed term in Equation 2 has the same form as Bock's (1972) NRM; but the difference is that the denominator is calculated only in terms of incorrect responses. Following Bock (1972), an arbitrary linear restriction is imposed for the incorrect response parameters as $\sum_{v=1}^{M} (\zeta_{iv} + \lambda_{iv}\theta_j) = 0$, implying that $\sum_{v=1}^{M} \zeta_{iv} = 0$ and $\sum_{v=1}^{M} \lambda_{iv} = 0$.

Taking into account a multilevel longitudinal structure, a multilevel longitudinal extension of the 2PL-NLM (hereafter, briefly called a multilevel longitudinal nested logit model [MLNLM]) is presented as follows. Assuming a two-level data structure in this illustration (e.g., students nested within teachers), the correct response probability is given as follows:

$$P(y_{ijt} = 1|\theta_{jgt}, \theta_{gt}) = \frac{\exp(\beta_{it} + \alpha_{1it}\theta_{jgt} + \alpha_{2it}\theta_{gt})}{1 + \exp(\beta_{it} + \alpha_{1it}\theta_{jgt} + \alpha_{2it}\theta_{gt})}, \tag{3}$$

where i and j are item and person indices, respectively, as in Equations 1 and 2. g is the index for a cluster (such as a teacher), t is the index for a time point, θ_{jgt} is an individual-level latent variable (Level 1), θ_{gt} is a cluster-level latent variable (Level 2), α_{1it} is an item discrimination parameter for θ_{jgt}, α_{2it} is an item discrimination parameter for θ_{gt}, and β_{it} is an item intercept parameter.[2] The cluster-level latent variable can be interpreted as the cluster mean of the individual-level latent variable in the case of $\alpha_{1it} = \alpha_{2it}$ for all items. The meaning of the latent variables over levels can differ otherwise.

Under the MLNLM (with a group invariance assumption), the probability that the examinee shows incorrect response category v is modeled as follows:

$$P(y_{ijt}=0, d_{ijvt}=1|\theta_{jgt}, \theta_{gt}) = P(y_{ijt}=0|\theta_{jgt}, \theta_{gt})P(d_{ijvt}=1|y_{ijt}=0, \theta_{jgt}, \theta_{gt})$$
$$= \left[1 - \frac{\exp(\beta_{it} + \alpha_{1it}\theta_{jgt} + \alpha_{2it}\theta_{gt})}{1 + \exp(\beta_{it} + \alpha_{1it}\theta_{jgt} + \alpha_{2it}\theta_{gt})}\right] \left[\frac{\exp(\zeta_{ivt} + \lambda_{1ivt}\theta_{jgt} + \lambda_{2ivt}\theta_{gt})}{\sum_{k=1}^{M}\exp(\zeta_{ikt} + \lambda_{1ikt}\theta_{jgt} + \lambda_{2ikt}\theta_{gt})}\right], \quad (4)$$

where v is the index for incorrect response categories, assuming group invariance as in Equation 3. λ_{1ivt} and λ_{2ivt} are the item discrimination parameters of a category v for θ_{jgt} and θ_{gt}, respectively, and ζ_{ivt} is an intercept category parameter. Similar to α_{1it} and α_{2it} in Equation 3, λ_{1ivt} and λ_{2ivt} can be identical if there is no cluster bias.[3]

Given there are T time points, the individual-level latent variables across time points, $\boldsymbol{\theta}_{jg} = [\theta_{jg1}, \theta_{jg2}, \ldots, \theta_{jgT}]'$, are assumed to follow a multivariate normal (MN) distribution, $\boldsymbol{\theta}_{jg} \sim MN(\boldsymbol{\mu}_{1(T\times1)}, \Sigma_{1(T\times T)})$. Likewise, the cluster-level latent variables across time points, $\boldsymbol{\theta}_{g} = [\theta_{g1}, \theta_{g2}, \ldots, \theta_{gT}]'$, are assumed to follow an MN distribution, $\boldsymbol{\theta}_{g} \sim MN(\boldsymbol{\mu}_{2(T\times1)}, \Sigma_{2(T\times T)})$. Meanwhile, $\Sigma_{1(T\times T)}$ and $\Sigma_{2(T\times T)}$ are unrestricted variance–covariance matrices. For model identification, $\boldsymbol{\mu}_{1(T\times1)}$ is set to $\mathbf{0}$ and the variances of $\Sigma_{1(T\times T)}$ are set to $\mathbf{1}$ with an assumption of cluster invariance.[4] When the item discrimination parameters are the same across levels (i.e., cluster bias does not exist), the variances of $\Sigma_{2(T\times T)}$ can be estimated. The variances of $\Sigma_{2(T\times T)}$ are also set to $\mathbf{1}$ when the discrimination parameters are different between Levels 1 and 2 (i.e., cluster bias exists).

Bayesian Estimation

To estimate the MLNLM, Bayesian estimation is used because of the high complexity of the models. WinBUGS 1.4.3 (Spiegelhalter, Thomas, Best, & Lunn, 2003) is used to implement a Markov chain Monte Carlo (MCMC) algorithm. The prior distributions of the model parameters are selected to resemble those used in previous studies (e.g., Bolt, Wollack, & Suh, 2012; Patz & Junker, 1999) and are presented here in WinBUGS specification. The prior distributions of the item parameters are $\alpha_{1it} \sim logNormal(0, 2)$, $\alpha_{2it} \sim logNormal(0, 2)$, $\beta_{it} \sim Normal(0, 0.5)$, $\lambda_{1ivt} \sim Normal(0, 0.5)$, $\lambda_{2ivt} \sim Normal(0, 0.5)$, and $\zeta_{ivt} \sim Normal(0, 0.5)$. Assuming the number of time points (T) is fixed at 2 in this application, the prior distribution of the individual-level latent variables follows a bivariate normal distribution (BN): $\boldsymbol{\theta}_{jg} = [\theta_{jg1}, \theta_{jg2}]' \sim BN(\mathbf{0}, \Sigma_{1(2\times2)})$, where the variances of $\Sigma_{1(2\times2)}$ are set to $\mathbf{1}$, while the correlation of $\Sigma_{1(2\times2)}$, $\rho_{\theta_{jg1}\theta_{jg2}} \sim Uniform(-1, 1)$. The prior distribution of the cluster-level latent variables is $\boldsymbol{\theta}_{g} = [\theta_{g1}, \theta_{g2}]' \sim BN(\boldsymbol{\mu}_{2(2\times1)}, \Sigma_{2(2\times2)})$, where the variances of $\Sigma_{2(2\times2)}$ are set to $\mathbf{1}$, when the item discrimination parameters are different between Levels 1 and 2, while the correlation of $\Sigma_{2(2\times2)}$, $\rho_{\theta_{g1}\theta_{g2}} \sim Uniform(-1, 1)$. When the item discrimination parameters are the same across all levels, the prior distribution of $\Sigma_{2(2\times2)}$ is as follows: $\Sigma_{2(2\times2)} \sim Wishart(\Omega_{2\times2}, 2)$, where $\Omega_{2\times2}$ is a unit matrix of size 2 and degrees of freedom 2 as the rank of $\boldsymbol{\theta}_{g}$. The hyperprior distribution on the mean is set at $\boldsymbol{\mu} \sim N(0, 1)$ for each time point. The convergence of the MCMC solution is evaluated with the Gelman and Rubin (1992) method with two chains.

Bayesian Model Fit

A relative fit criterion, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), is used as a tool for comparing the fit of the models being considered. The DIC is calculated as $DIC = \bar{D}(\boldsymbol{\vartheta}) + p_D$, where $\boldsymbol{\vartheta}$ is a set of parameters, $\bar{D}(\boldsymbol{\vartheta})$ is the posterior mean of the deviance (i.e., a Bayesian measure of fit), and p_D is called the effective

number of parameters (i.e., a Bayesian measure of complexity). p_D is defined as $p_D = \bar{D}(\boldsymbol{\vartheta}) - D(\hat{\boldsymbol{\vartheta}})$, where $\hat{\boldsymbol{\vartheta}}$ is the posterior estimates of the parameters, and $D(\hat{\boldsymbol{\vartheta}})$ is the deviance obtained from the posterior estimates of the parameters. The DIC can be easily computed by obtaining the two deviance values, $\bar{D}(\boldsymbol{\vartheta})$ and $D(\hat{\boldsymbol{\vartheta}})$, as the outcomes of the WinBUGS runs. A smaller DIC indicates a better fit, and a difference of less than 5 or 10 units between the models does not provide sufficient evidence for favoring one model over another (Spiegelhalter et al., 2003).

Empirical Illustration: Effects of Enhanced Anchored Instruction

Study Design, Samples, and Measures

This study analyzed data obtained via a host study funded by the Institute of Education Sciences (U.S. Department of Education). The goal of the host study was to evaluate the efficacy of enhanced anchored instruction (EAI), which was designed by Bottge and his colleagues. For more details on the study, the full report can be found in Bottge, Ma, Gassaway, Toland, et al. (2014). The design of the study is a pretest–posttest, cluster randomized trial to compare the effects of two instructional conditions (EAI vs. business as usual [BAU]) on students' computation and problem-solving skills.

A total of 24 urban and rural middle schools in the Southeast participated in the study. Schools were randomly assigned to EAI and BAU (12 EAI and 12 BAU schools). Each school had one math classroom except one school, which had two math classrooms. A total of 471 students were in the initial sample. Of these students, 25 students did not respond to all items on the pretest or posttest. Consequently, 232 BAU students and 214 EAI students remained in the final sample. Twenty-nine percent of the BAU students had math difficulties (MD), while 26% of the EAI students had MD. The smallest number of students analyzed for each classroom (i.e., teacher) was seven, and the largest was 28. The data showed a two-level structure. That is, students were nested in teachers. Based on the chi-square tests of equal proportions, the students were comparable between the EAI and BAU groups in terms of gender, ethnicity, subsidized lunch, and disability area, and the teachers were also comparable in gender, ethnicity, and education level (see Bottge, Ma, Gassaway, Toland, et al., 2014, for the results).

A researcher-developed test called the Fraction and Computation Test (FCT) was used to measure students' ability to add and subtract factions. The tests that had 20 items were administered twice, as a pretest (Time 1) and a posttest (Time 2). The 20 items possess different item attributes[5]: Students were asked to add and subtract fractions with *like* denominators (e.g., $1/5 + 2/5$) or *unlike* denominators (e.g., $5/6 - 1/3$), *simple* fractions (e.g., $5/8 + 1/4$) or *mixed* numbers (e.g., $4\,1/16 + 1/8 + 1/2$), and two stacks (e.g., $3/4 / 1/2$) or three stacks (i.e., one more stack in the two-stack example). For the specified attributes for each item, see Table 2.

For 18 of the 20 items, students could earn 0, 1, or 2 points. For the other two items, students could get 3 points if they simplified the correct answers (i.e., reduced the fraction to a simple term). However, less than 1% of the students in the sample received partial scores on any of the items. In this study, partial scores were considered incorrect responses. Interrater agreement was 99% on the pretest and 97% on the posttest. Raters also identified and coded the primary error types the students made for each incorrect item. A total of 11 error codes[6] were generated and labeled as shown in Appendix A as an online supplement. There were no missing item responses in the data set.

Table 1. Model Comparison Results.

Model	DIC
Longitudinal NLM	30,110
MLNLM with cluster invariance	30,010
MLNLM without cluster invariance	29,360

Note. DIC = deviance information criterion; NLM = nested logit model; MLNLM = multilevel longitudinal nested logit model.

Model Comparisons

Using the FCT data, models were compared in terms of model fit with the DIC.[7] To analyze the data with 11 error types (categories) and one correct answer, three models[8] were fitted. The first model was a longitudinal NLM without a multilevel structure (i.e., Equations 3 and 4 without the term θ_{gt}). The second model was the MLNLM with an assumption of cluster invariance (i.e., Equations 3 and 4 with $\alpha_{1it} = \alpha_{2it}$ and $\lambda_{1ivt} = \lambda_{2ivt}$). The third model was the MLNLM without an assumption of cluster invariance (i.e., Equations 3 and 4 with $\alpha_{1it} \neq \alpha_{2it}$ and $\lambda_{1ivt} \neq \lambda_{2ivt}$). Based on the convergence check results, a conservative burn-in of 6,000 iterations was used, followed by 4,000 postburn-in iterations for all analyses. Table 1 provides the model comparison results. The MLNLM without the cluster invariance assumption appears to be the best-fitting model (with a DIC value of 29,360) among the three models.[9] Therefore, the MLNLM with cluster bias was chosen as the base model for subsequent analyses in this study. As explained earlier, when the cluster invariance assumption does not hold (i.e., $\alpha_{1it} \neq \alpha_{2it}$ and $\lambda_{1ivt} \neq \lambda_{2ivt}$), the cluster-level latent variable (θ_{gt}) is not the cluster mean of the individual-level latent variable (θ_{jgt}) that can be labeled as students' ability to add and subtract factions in this example. Because an instructional intervention EAI was implemented at the cluster (i.e., teacher) level, the cluster-level latent variable (θ_{gt}) could be interpreted as the EAI-related student ability, which varied across teachers (i.e., EAI effect for EAI students vs. no EAI effect for BAU students).

Sensitivity Analysis

Before inferences were made about intervention effects with the FCT data, group and time invariance assumptions were investigated as a preliminary analysis under the MLNLM with cluster bias (the model selected from the model comparison study). The detailed procedures are described in Appendix B as an online supplement. As a result of the measurement invariance study, no group bias was found at the cluster- and individual-group levels. For the time invariance study, bias in the intercept parameter for the correct response (β_{it}) was found between the two time points. To assess the impact of having the bias in the intercept on the results of the latent variables (θs), a sensitivity analysis was conducted. The goal of the analysis was to examine the consequences of ignoring the bias by fitting the strong invariance model versus the weak time invariance model for the correct response (i.e., bias in the intercept). The sensitivity analysis was performed in terms of the estimates (i.e., posterior means from WinBUGS outputs) of the individual- and cluster-level latent variables (θ_{jgt} and θ_{gt}) obtained from both models.[10]

By fitting each model, two individual-level latent variables, θ_{jg1} for the pretest (Time 1) and θ_{jg2} for the posttest (Time 2), and two cluster-level latent variables, θ_{g1} for the pretest and θ_{g2} for the posttest, were estimated as the outcomes of the model. First, correlations were computed between the weak and strong invariance models for each latent variable type. The correlation

coefficients were sufficiently high: .99, .99, .99, and .98 for $\hat{\theta}_{jg1}$,[11] $\hat{\theta}_{jg2}$, $\hat{\theta}_{g1}$, and $\hat{\theta}_{g2}$, respectively. Moreover, within each model, difference scores were calculated as $\hat{\theta}_{jg2} - \hat{\theta}_{jg1}$ (i.e., posttest − pretest) for the individual-level latent variable and $\hat{\theta}_{g2} - \hat{\theta}_{g1}$ (i.e., posttest − pretest) for the cluster-level latent variable. The means and variances for each difference score were very similar across the models.[12] Finally, two *t* tests were conducted using the means and variances of the difference scores from the two models: one test for the difference between the two models using the two means of $\hat{\theta}_{jg2} - \hat{\theta}_{jg1}$ (i.e., an independent samples *t* test with the null hypothesis saying that the mean of the difference score from the weak invariance model is the same as the mean from the strong invariance model) and another test for the difference between the two models using the two means of $\hat{\theta}_{g2} - \hat{\theta}_{g1}$. For both tests, the *t*-test results—$t = 0.58$ ($p = .56$, $df = 890$) for the individual level and $t = 0.90$ ($p = .38$, $df = 48$) for the cluster level—showed that the means of the difference scores from the weak invariance model were not significantly different from the means from the strong model at both levels.

These results suggested that although bias in the intercept parameter for the correct response (β_{it}) was found between the two time points, fitting the strong invariance model (i.e., ignoring the bias) did not yield substantially different results from fitting the weak time invariance model for the correct response. In addition, Verhagen and Fox (2013) noted that measurement invariance should be at least weakly invariant to make comparisons across groups and to measure changes across time points. Therefore, it would be reasonable to consider one-group MLNLM (the strong invariance model) with cluster bias as the final measurement model for subsequent analyses.

Instructional Effects With the Results of the MLNLM

In this section, the instructional effects on the probabilities of correct response and each error type were examined using the (one-group) MLNLM with cluster bias, as the final model based on the model comparisons and the measurement invariance investigations. The substantive research questions in this section were as follows: (a) What are the differential effects of BAU and EAI on the correct response probability and error patterns? (b) What are the differential effects of BAU and EAI within the non-MD and MD groups on the correct response probability and error patterns?

Table 2 provides only the item parameter estimates (posterior means) and a 95% credibility interval (CI) for the correct response ($\hat{\alpha}_{1i}$, $\hat{\alpha}_{2i}$, and $\hat{\beta}_i$) due to page limits. The item parameter estimates for the error responses ($\hat{\lambda}_{1iv}$, $\hat{\lambda}_{2iv}$, and $\hat{\zeta}_{iv}$) are shown in Appendix C as an online supplement. Items that had a *like* denominator were less discriminating and easier than items that had an *unlike* denominator. No particular pattern for the other item attributes was found. The means and standard deviations (*SD*s) of the estimated latent variables ($\hat{\theta}_{jgt}$ and $\hat{\theta}_{gt}$) at each time point were obtained—$\bar{\theta}_{jg1} = -0.83$, $\bar{\theta}_{jg2} = 3.17$, $\bar{\theta}_{g1} = -3.31$, $\bar{\theta}_{g2} = -3.21$; $SD(\hat{\theta}_{jg1}) = 6.49$, $SD(\hat{\theta}_{jg2}) = 6.33$, $SD(\hat{\theta}_{g1}) = 0.69$, and $SD(\hat{\theta}_{g2}) = 0.90$. The difference in the mean of each latent variable between the posttest (Time 2) and the pretest (Time 1) was calculated as follows: $\bar{\theta}_{jg2} - \bar{\theta}_{jg1} = 3.17 - (-0.83) = 4.00$ and $\bar{\theta}_{g2} - \bar{\theta}_{g1} = -3.21 - (-3.31) = 0.10$, implying there were increases in the individual-level latent variable (i.e., students' ability to add and subtract fractions) and in the cluster-level latent variable (i.e., EAI-related students' ability), on average, after the intervention, respectively.

To answer the research questions, the correct response probability and the probability of each error type were calculated using the item and person parameter estimates of the MLNLM. Regarding the differential effects of BAU versus EAI, the differences in the probabilities between the BAU and EAI groups were calculated across 12 response categories (one correct response and 11 error types) for 20 items at each time point. The difference was calculated as

Table 2. Correct Response Item Parameter Estimates and 95% Credibility Intervals under the MLNLM with Cluster Bias.

	Attributes						
Item	Operation	Denominator	Type	Stacks	$\hat{\alpha}_{1i}$ (95% CI)	$\hat{\alpha}_{2i}$ (95% CI)	$\hat{\beta}_i$ (95% CI)
1	Addition	like	simple	2	0.26 [0.21, 0.30]	0.27 [0.14, 0.43]	2.68 [2.19, 3.22]
2	Addition	like	simple	2	0.29 [0.24, 0.34]	0.23 [0.12, 0.37]	2.83 [2.36, 3.34]
3	Addition	unlike	simple	2	0.80 [0.67, 0.94]	0.99 [0.64, 1.32]	0.92 [−0.03, 1.94]
4	Addition	unlike	simple	2	0.67 [0.57, 0.79]	0.95 [0.66, 1.24]	1.06 [0.25, 1.99]
5	Addition	unlike	simple	2	0.71 [0.59, 0.84]	1.03 [0.72, 1.37]	0.32 [−0.59, 1.22]
6	Addition	unlike	simple	2	0.93 [0.77, 1.12]	1.06 [0.70, 1.45]	0.24 [−0.78, 1.41]
7	Addition	unlike	mixed	2	0.71 [0.60, 0.83]	0.84 [0.58, 1.17]	0.01 [−0.84, 0.98]
8	Addition	unlike	mixed	2	0.59 [0.50, 0.69]	0.68 [0.42, 0.96]	−0.41 [−1.24, 0.47]
9	Addition	unlike	mixed	2	0.77 [0.65, 0.91]	1.05 [0.72, 1.45]	0.00 [−0.86, 0.94]
10	Addition	unlike	mixed	2	0.68 [0.57, 0.81]	1.14 [0.82, 1.48]	0.44 [−0.37, 1.37]
11	Addition	unlike	simple	3	0.60 [0.52, 0.71]	0.85 [0.61, 1.11]	0.70 [−0.13, 1.52]
12	Addition	unlike	simple	3	0.69 [0.58, 0.82]	1.01 [0.70, 1.34]	0.43 [−0.52, 1.34]
13	Addition	unlike	mixed	3	0.67 [0.56, 0.79]	0.82 [0.55, 1.08]	−0.73 [−1.50, 0.00]
14	Addition	unlike	mixed	3	0.58 [0.48, 0.69]	0.96 [0.67, 1.25]	−0.32 [−1.14, 0.43]
15	Subtraction	like	simple	2	0.19 [0.15, 0.23]	0.21 [0.11, 0.34]	2.28 [1.90, 2.79]
16	Subtraction	unlike	simple	2	0.64 [0.54, 0.75]	0.85 [0.56, 1.11]	0.05 [−0.74, 0.74]
17	Subtraction	like	mixed	2	0.16 [0.13, 0.19]	0.19 [0.11, 0.30]	1.16 [0.85, 1.58]
18	Subtraction	unlike	mixed	2	0.29 [0.24, 0.35]	0.81 [0.58, 1.08]	−0.08 [−0.77, 0.77]
19	Subtraction	unlike	mixed	2	0.53 [0.45, 0.63]	0.87 [0.58, 1.14]	−0.21 [−1.01, 0.57]
20	Subtraction	unlike	mixed	2	0.32 [0.26, 0.39]	0.79 [0.51, 1.09]	−1.23 [−2.08, −0.40]

Note. CI = credibility interval; MLNLM = multilevel longitudinal nested logit model.

the probability in BAU minus the probability in EAI. For easier representation, the differences in probabilities between BAU and EAI were plotted for each item. Positive values were expected for the differences for the error types and negative values for the correct responses on the posttest (Time 2) in the presence of instructional effects. The differences were expected to be close to zero on the pretest (Time 1), in which instruction had not been introduced.

After the plots of all items were inspected, the anticipated pattern was observed. An interesting finding was that the pattern was more obvious in a certain attribute over the other attributes. For example, Figure 1a shows plots for *like* items and *unlike* items. The differences were averaged over *like* items (four items) and the *unlike* items (16 items) separately and are displayed in the figure. A common pattern across the two plots is that the differences in the probability between BAU and EAI on the pretest were minimal (see the line for Time 1), but there were instructional effects between EAI and BAU students in making the errors and endorsing the items (see the line for Time 2). In particular, the difference in the probability of the *Combining* (C: Student combines numerators together and denominators together) error was the highest among the 11 error types, implying that the EAI students noticeably reduced the C error from the pretest compared with the BAU students. Furthermore, the error probability for the *Other* (O: Student makes errors other than those labeled) error was improved (i.e., reduced) for the EAI students. For the correct response probability, the probability increased after instruction. These patterns were more apparent in the *unlike* items than in the *like* items. For the *unlike* items, the probability of the C error for the EAI students was 0.15 lower than that for BAU students, and the EAI students showed a 0.25 higher probability of solving the item correctly than the BAU students after instruction. Regarding the addition versus subtraction items, the addition items showed a clearer pattern than the subtraction

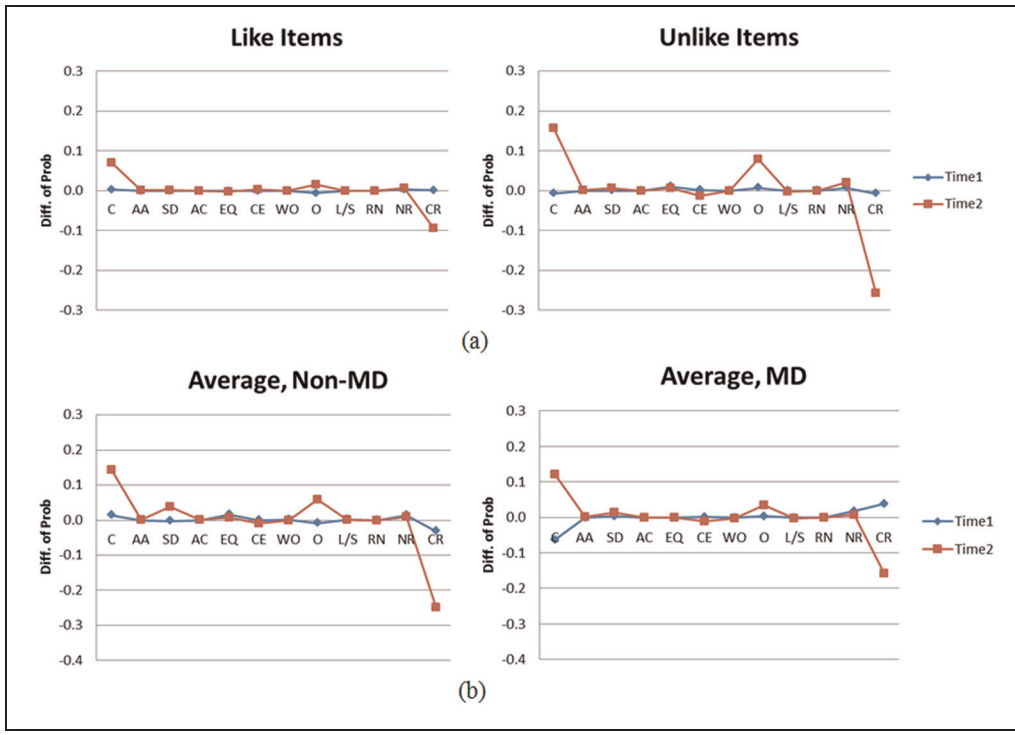

Figure 1. (a) Average differences (the probability in BAU minus the probability in EAI) in the response probabilities for like items (left) and unlike items (right); (b) average differences (the probability in BAU minus the probability in EAI) across all items within non-MD (left) and MD (right).

Note. BAU = business as usual; EAI = enhanced anchored instruction; MD = math difficulties; C = combining; AA = add all; SD = select denominator; AC = adding components; EQ = equivalent fraction error; CE = computation error; WO = wrong operation; O = other; L/S = large–small; RN = renaming; NR = no response; CR = correct response.

items. A previous study showed that fewer *C* errors by students in the EAI group contributed to significant improvements in adding-related items overall (Bottge, Ma, Gassaway, Butler, et al., 2014). No difference was observed between the simple and mixed items. Finally, compared with two-stack items, three-stack items presented results that were more similar to those for the *unlike* items and addition items.

The differential effect of BAU and EAI on the correct response probability and error probabilities was examined within the non-MD versus MD groups. The differences were averaged across all 20 items and plotted. Figure 1b shows the average differences within each group. Similar patterns found in Figure 1a were observed. Interestingly, in the non-MD group, the EAI students made fewer *Select Denominator* (SD: Student selects one of the denominators listed in the problem and makes no attempt to make an equivalent fraction) errors than the BAU students. In the MD group, the EAI students made more C errors on the pretest than the BAU students. However, the positive effect of improving performance on C errors was greater for the MD group than for the non-MD group. In other words, the absolute difference in the probability of C errors between Time 2 and Time 1 was higher in the MD group than in the non-MD group. The instructional effect on the correct response probability was about 0.1 higher in the non-MD group than in the MD group. Regarding item attributes, the addition, *unlike*, and three-stack

items showed more evident patterns, as in Figure 1b, compared with the subtraction, *like*, and two-stack items, respectively. No difference was found between simple and mixed items.

Parameter Recovery Study

To evaluate the parameter recovery of the MLNLM with cluster bias chosen as the final measurement model, a simulation study was conducted. First, using the parameter estimates obtained from the real data analysis as the true parameters, item responses for 446 examinees nested within 25 clusters and 20 items (as found in the empirical study) were generated under the MLNLM. Second, to investigate the effect of increasing the sample size on the parameter recovery, the cluster size was increased from 25 to 50, and the number of examinees within each cluster was also doubled, resulting in 892 examinees in total. Thus, the two sample sizes were simulated (small vs. large in relative terms) under the MLNLM. For each sample size condition, 30 replications were simulated. Finally, the effect of ignoring cluster membership on the parameter estimation was examined additionally by fitting a single-level NLM instead of the MLNLM. This was done only in the large sample condition. In all simulation conditions, the true model was the MLNLM.

As in the real data analysis, chains were simulated to 10,000 iterations with a burn-in of the first 6,000 iterations. The prior distributions were identical to those used in the real data estimation. The convergence of the MCMC solution was evaluated with the Gelman and Rubin (1992) method with two chains. Recovery results were summarized in terms of relative bias and the root mean square error (RMSE) of the item parameter estimates relative to the true parameters. For example, for α_{1i}, the relative bias and the RMSE were obtained with $\sum_{r=1}^{30}[(\hat{\alpha}_{1i}-\alpha_{1i})/\alpha_{1i}]/30$ and $\sqrt{\sum_{r=1}^{30}(\hat{\alpha}_{1i}-\alpha_{1i})^2/30}$, respectively. For the sake of interpretation, the relative biases and the RMSEs[13] for individual error categories were later collapsed across the error response categories, as well as items, to create an average value for each category parameter type. The relative biases and the RMSEs for the correct response category parameters were averaged across items. The relative bias implies the accuracy of the parameter estimates, and the RMSE combines the parameter bias and precision into an overall measure of accuracy. The authors considered that values of relative bias of larger than $|0.2|$ were unacceptable (e.g., Forero & Maydeu-Olivares, 2009).

Table 3 reports the recovery results for the six item parameters, α_{1i}, α_{2i}, β_i, λ_{1iv}, λ_{2iv}, and ζ_{iv}. When the MLNLM was fitted, the relative biases for all item parameters were not larger than $|0.2|$. Based on the RMSE values, the parameter recovery of the MLNLM with the large sample condition improved slightly in some parameters (i.e., α_{1i}, α_{2i}, λ_{1iv}, and ζ_{iv}) compared with the recovery with the small sample condition. In other parameters, the RMSE stayed the same (λ_{2iv}) or unexpectedly increased (β_i). Given the large number of categories (12), the RMSEs for most category parameters appeared comparable with those seen in previous studies under the 2PL (e.g., Patz & Junker, 1999), the 2PL-NLM (Bolt et al., 2012; Suh & Bolt, 2010), and the NRM (Wollack, Bolt, Cohen, & Lee, 2002). For the intercept parameters, β_i and ζ_{iv}, relatively high RMSEs were observed; however, the results can be reasonably compared to the study by Wollack et al. (2002), in which two sample sizes of 300 and 500 examinees and three test lengths (10, 20, and 30 items) were used. Four response categories, including the correct response, were considered, and MML and MCMC estimation methods were implemented by Wollack et al. (2002). Using a 20-item condition with the MCMC estimation, Wollack et al. (2002) reported that the average RMSEs for the slope parameters across item categories under the NRM ranged from 0.14 to 0.29, and the average intercept RMSEs ranged from 0.13 to 0.46 depending on ability levels.

Table 3. Item Parameter Recovery Results.

Model	α_{1i}	α_{2i}	β_i	λ_{1iv}	λ_{2iv}	ζ_{iv}
Relative bias						
MLNLM_SM	0.02	0.12	−0.02	−0.13	−0.11	−0.20
MLNLM_LG	−0.01	−0.06	0.17	−0.18	0.17	−0.06
NLM_LG	0.02		−1.10	−0.14		−0.43
RMSE						
MLNLM_SM	0.05	0.13	0.35	0.15	0.24	0.41
MLNLM_LG	0.03	0.09	0.45	0.12	0.24	0.34
NLM_LG	0.03		0.99	0.11		1.00

Note. MLNLM = multilevel longitudinal nested logit model; MLNLM_SM = fitting the MLNLM to the small sample size condition; MLNLM_LG = fitting the MLNLM to the large sample size condition; NLM_LG = fitting a single-level NLM to the large sample size condition; NLM = nested logit model; RMSE = root mean square error.

Finally, when the single-level NLM was fitted to the large sample size condition (i.e., when cluster membership was ignored), the relative biases for the intercept parameters in the correct and incorrect response categories (i.e., β_i and ζ_{iv}) were substantially larger than $|0.2|$. In addition, the RMSEs for β_i and ζ_{iv} with the single-level NLM were more than twice those with the MLNLM, yielding values of 0.99 and 1.00. However, ignoring the multilevel structure did not seem to affect the estimation of the slope parameters.

Discussion and Conclusion

The MLNLM was presented to analyze nominal responses under a pretest–posttest, cluster randomized trial design, which is one of the commonest designs used in educational intervention studies. Using the real data analysis, this article illustrates the application of the model to a model comparison study for selecting the best-fitting measurement model and for examining cluster bias (and measurement invariance over time points and between group variables at the individual and cluster levels, respectively, in the online supplement). Based on the model comparison and measurement invariance studies, the (one-group) MLNLM with cluster bias was chosen as the final model for examining the instructional effects of BAU and EAI on the correct response and error response probabilities. A parameter recovery study was conducted by mimicking the conditions of the real data set as the simulation design. The effect of increasing the sample size was also examined by doubling the sample size in the real data set. The accuracy of the item parameter estimates under the MLNLM was comparable with those found in previous parameter recovery studies given the large number of response categories. As the sample size increased, the parameter recovery improved slightly in some parameters (α_{1i}, α_{2i}, λ_{1iv}, and ζ_{iv}) but not in others (β_i and λ_{2iv}). In this regard, further research on the parameter recovery of the MLNLM might be needed in relation to changes in the sample size and the number of response categories. Finally, ignoring cluster membership adversely affected the estimation of the intercept parameters.

Future studies with the MLNLM can address various issues. First, the DIC was used to examine the model's fit for the model comparison and measurement invariance checks as used in other multilevel IRT studies (e.g., Cho & Bottge, 2015; Fox, 2010). However, the performance of the DIC has not yet been examined under the current application of the MLNLM in terms of identifying the correct model and detecting the sources of measurement variance. Particularly when the sample size and the number of clusters are small under hierarchical

models, the asymptotic properties of the DIC and robustness to nonnormal distributions can be impaired. The DIC also tends to choose overfitted models (i.e., more complex models). Therefore, practitioners need to be cautious when they choose the DIC as a relative model fit criterion. Detailed discussions on the limitations and cautions of using the DIC have been documented by Spiegelhalter, Best, Carlin, and van der Linde (2014). In this regard, it would be valuable to evaluate the performance of the DIC via a simulation study in terms of model comparisons in the presence of hierarchical data structures, as well as Type I errors and power for a measurement invariance study under the MLNLM or its variants.

Second, only the MLNLM with cluster bias was examined in the recovery study by mimicking real data testing conditions. Additional investigations under the MLNLM without cluster bias and/or more extensive simulation studies under various testing conditions, including different sample sizes and test lengths, should be considered for a future study. In addition, it would be worthwhile to examine the consequences of ignoring cluster bias (i.e., assuming cluster invariance) in measuring individual- and cluster-level latent variables and changes over time.

Third, cluster bias was examined at the test level, rather than at the individual-item level. However, there might be cluster bias for some, but not all, test items. This is referred to as partial cluster invariance (Jak et al., 2013). Considering partial cluster invariance after examining test-level cluster bias might help researchers explore why cluster bias presents based on item and cluster information. With the assumption of cluster bias under the MLNLM, the individual- and cluster-level latent variables cannot be interpreted in the same way across all levels. The interpretation of latent variables at each level would be of practical interest based on the individual-item level analysis when item content information is available.

Fourth, this study followed previous studies and conventional practices in choosing priors. However, the effect of priors on parameter estimation (such as the variances of individual- and cluster-level latent variables) was not examined thoroughly although some preliminary analyses were done before using the selected priors. Therefore, conducting a sensitivity analysis to choose proper priors under various simulation conditions would be valuable.

Fifth, as illustrated in the real data analysis, the effects of instruction (BAU vs. EAI) and group memberships (non-MD and MD groups) were examined using the parameter estimates from the MLNLM. However, the MLNLM can be extended by incorporating the instruction and group variables in the model as person covariates to explain the individual- and cluster-level latent variables as in explanatory IRT models (De Boeck & Wilson, 2004), which will be a more complex model than the MLNLM but can provide direct estimates of the instructional and/or group effects on the latent variables.

Finally, there exist some concerns about the use of hierarchical models. Compared with the HLM, the MLNLM may perform adequately when cluster size, the number of clusters, and intraclass correlations are large because there are more parameters in the MLNLM than in the HLM. Therefore, further research on these design factors regarding hierarchical structures for the MLNLM is still needed.

The main goal of this study was to present and illustrate the MLNLM to analyze the correct response and error types as an IRT model for detecting intervention effects in multilevel longitudinal intervention data. A major advantage of the MLNLM is its ability to present the results based on the correct response and error probability comparisons, which together can provide practitioners with a detailed picture of the strengths and weaknesses of the intervention.

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: The data used in the article were collected with the following support: the U.S. Department of Education, Institute of Education Sciences, PR Number H324A090179. Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the supporting agency.

Notes

1. Multilevel item response theory (IRT) models assuming the same item discriminations over levels can be considered variance component factor models (Rabe-Hesketh, Skrondal, & Pickles, 2004).
2. For the item parameters, group invariance is assumed in terms of notation. If the group invariance assumption is violated, then the g index needs to be included as a subscript in each item parameter, implying the parameters are freely estimated across the groups being considered. In addition, α_{1it} and α_{2it} can be identical if there is no cluster bias—that is, $\exp(\beta_{it} + \alpha_{1it}\theta_{jgt} + \alpha_{2it}\theta_{gt})$ is reduced to $\exp[\beta_{it} + \alpha_{1it}(\theta_{jgt} + \theta_{gt})]$. Finally, α_{1it}, α_{2it}, and β_{it} in Equation 3 can be reduced to α_{1i}, α_{2i}, and β_i, respectively, with the assumption of time invariance.
3. λ_{1ivt}, λ_{2ivt}, and ζ_{ivt} can be reduced to λ_{1iv}, λ_{2iv}, and ζ_{iv}, respectively, with the assumption of time invariance. The same constraints pertaining to Equation 2 are applied—namely, $\sum_{v=1}^{M} \zeta_{ivt} = 0$, $\sum_{v=1}^{M} \lambda_{1ivt} = 0$, and $\sum_{v=1}^{M} \lambda_{2ivt} = 0$ (i.e., sum-to-zero constraints for each time point t) are imposed.
4. When the individual-level latent variable and cluster-level latent variable are orthogonal and one latent variable is considered at each level for each time point, additional constraints on the individual- and cluster-level item discriminations are not necessary (Longford & Muthén, 1992, pp. 583-584).
5. As a preliminary analysis, exploratory factor analyses were conducted at each time point in Mplus version 7.11 (L. K. Muthén & Muthén, 1998-2015). Although the 20 items can be grouped based on the item features (e.g., like vs. unlike items), the one-factor solution provided a good fit to the data.
6. Error type coding assumed that all items have the same 11 error types.
7. Before the deviance information criterion (DIC) was applied for model comparisons, the posterior distributions of model parameters were examined. The distributions were symmetrical; that is, the mean and the median showed similar values.
8. For this model comparison analysis, time invariance was presumed across all the models. As shown in the next section, a preliminary analysis was conducted in terms of measurement invariance checks over the time points and the individual- and cluster-level groups.
9. The multilevel longitudinal nested logit model (MLNLM) without the cluster invariance assumption (i.e., MLNLM with cluster bias) can be fitted for the correct response category only. That is, α_{1it} and α_{2it} are freely estimated, while λ_{1ivt} and λ_{2ivt} are assumed to be identical. The opposite case can be modeled, too. These two models were fitted, and their DIC values were greater than the DIC value (29,360) of the best-fitting model (i.e., the MLNLM with cluster bias for the correct response and incorrect response categories).
10. Note that the only difference between the two models is that two different intercept parameters (β_{it}; one for each time point) can be estimated across time points in the weak invariance model, whereas the same intercept parameter (β_i) between the two time points is obtained in the strong invariance model.
11. Hereafter, the hat over each parameter represents the estimate of the true parameter.
12. The means and variances of the individual-level difference scores were 3.57 and 34.74 for the weak model, and 3.34 and 37.77 for the strong model, respectively, and the means and variances of the cluster-level difference scores were 0.27 and 0.46 for the weak model and 0.24 and 0.49 for the strong model, respectively.
13. Note that the biases for the error categories are naturally forced to zero because of the sum-to-zero constraints applied in Equation 4. In addition, the biases and the root mean square errors (RMSEs) for the correct response category parameters showed similar patterns. Therefore, we reported the relative bias and the RMSE only.

Supplemental Material

Supplemental material for this article is available online.

References

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

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bolt, D. M., Wollack, J. A., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, *77*, 339-357.

Bottge, B. A., Ma, X., Gassaway, L., Butler, M., & Toland, M. D. (2014). Detecting and correcting fractions computation error patterns. *Exceptional Children*, *80*, 236-254.

Bottge, B. A., Ma, X., Gassaway, L., Toland, M. D., Butler, M., & Cho, S.-J. (2014). Effects of blended instructional models on math performance. *Exceptional Children*, *80*, 423-437.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581-612.

Charalambous, C. Y., & Pitta-Pantazi, D. (2007). Drawing on a theoretical model to study students' understandings of fractions. *Educational Studies in Mathematics*, *64*, 293-316. doi:10.1007/s10649-006-9036-2

Cho, S.-J., & Bottge, B. A. (2015). Multilevel multidimensional item response model with a multilevel latent covariate. *British Journal of Mathematical and Statistical Psychology*, *68*, 410-433.

Clarke, D. M., & Roche, A. (2009). Students' fraction comparison strategies as a window into robust understanding and possible pointers for instruction. *Educational Studies in Mathematics*, *72*, 127-138. doi:10.1007/s10649-009-9198-9

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*, 300-315.

De Boeck, P., & Wilson, M. (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.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275-299.

Fox, J.-P. (2010). *Bayesian item response modeling*. New York, NY: Springer.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271-288.

Fox, J.-P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169-191.

Fox, J.-P., & Glas, C. A. W. (2016). Multilevel item response models with covariates and multiple groups. In W. J. van der Linden (Ed.), *Handbook of item response theory, models, statistical tools, and applications* (Vol. 1., pp. 407-420). Boca Raton, FL: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 265-282.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

Kim, J.-S. (2006). Using the distractor categories of multiple-choice items to improve IRT linking. *Journal of Educational Measurement*, *43*, 193-213.

Lee, W.-y., & Cho, S.-J. (2015, April). *Detecting cluster bias in a multilevel item response model*. Annual Meeting of National Council on Measurement in Education, Chicago, IL.

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, *57*, 581-597.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 57-74). San Diego, CA: Academic Press.

Muthén, B. (1990). *Mean and covariance structure analysis of hierarchical data*. Los Angeles: University of California, Los Angeles.

Muthén, B., & Asparouhov, T. (2016). Multidimensional, multilevel, and multi-time point item response modeling. In W.J. van der Linden (Ed.), *Handbook of item response theory, models, statistical tools, and applications* (Vol. 1. pp. 527-540). Boca Raton, FL: Chapman & Hall.

Muthén, L. K., & Muthén, B. O. (1998-2015). Mplus [Computer program]. Los Angeles, CA: Author.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167-190.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173-185.

Raudenbush, S. W., & Bryk, A. G. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

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*, *64*, 583-616.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, *76*, 485-493.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, *75*, 454-473.

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*, 188-205.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577.

Verhagen, J., & Fox, J.-P. (2013). Longitudinal measurement in health-related surveys: A Bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine*, *32*, 2988-3005.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318-336.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *26*, 339-352.