Check for updates

Article

Journal of Early Adolescence
2017, Vol. 37(1) 85–128
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0272431616642328
jea.sagepub.com
SAGE

Introduction to Multilevel Item Response Theory Analysis: Descriptive and Explanatory Models

Isabella Sulis[1] and Michael D. Toland[2]

Abstract

Item response theory (IRT) models are the main psychometric approach for the development, evaluation, and refinement of multi-item instruments and scaling of latent traits, whereas multilevel models are the primary statistical method when considering the dependence between person responses when primary units (e.g., students) are nested within clusters (e.g., classes). This article introduces multilevel IRT (MLIRT) modeling, and provides the basic information to conduct, interpret, and report results based on an analysis using MLIRT modeling. The procedures are demonstrated using a sample data set based on the National Institute for the Evaluation of School System survey completed in Italy by fifth-grade students nested in classrooms to assess math achievement. The data and command files (Stata, M*plus*, flexMIRT) needed to reproduce all analyses and plots in this article are available as supplemental online materials at http://jea.sagepub.com/supplemental.

Keywords

multilevel modeling, item response theory, Stata, M*plus*, flexMIRT

[1]University of Cagliari, Italy
[2]University of Kentucky, Lexington, USA

Corresponding Author:
Isabella Sulis, Department of Social Sciences and Institution, University of Cagliari, Viale. S. Ignazio 78, Cagliari 09123, Italy.
Email: isulis@unica.it

In early adolescence research, a wide range of data structures and research designs generate dependence between observations that can be straightforwardly modeled by translating the data structure and the research questions into a multilevel framework (Goldstein & Thomas, 1996). Typical examples encountered in early adolescence include studies where researchers are interested in jointly modeling repeated measurements on the same adolescent over time, students nested within classes or schools, early adolescents belonging to the same youth club or sports team, and adolescents within peer groups. Multilevel modeling (MLM) takes into account the degree of dependency among respondents and has become the championed statistical method for such designs (Hox, 2010; Peugh, 2014; Raudenbush & Bryk, 2002). Also, early adolescence researchers commonly include variables in research that cannot be directly observed or measured without error. Many of these variables include psychological constructs (e.g., motivation, emotional intelligence), aptitudes, or more generally cognitive traits that are indirectly measured by means of a multi-item instrument. In order to properly model the nonlinear relationship that exists between item response behavior and the latent trait of interest, researchers can use item response theory (IRT; Birnbaum, 1968; Lord, 1952) or Rasch (1960).

Naturally, it is a common phenomenon in early adolescence studies to observe situations where a nested data structure exists and data on the primary response variable is derived from a multi-item instrument. To appropriately analyze the multi-item responses within a nested data structure (e.g., students within classes), a multilevel IRT model can be specified by allowing multiple item responses to be a function of the latent traits at both the student and class levels in a regression analysis (Fox, 2007, p. 2). This model can be specified using software that has the routine for fitting generalized linear and nonlinear mixed models (see De Boeck & Wilson, 2004; Fox, 2007; Jeon & Rabe-Hesketh, 2012; Skrondal & Rabe-Hesketh, 2004). Embedding IRT models within an MLM framework allows us to derive scores for latent traits on a metric or interval scale at different levels of analysis (e.g., students, classes, and schools) while accounting for measurement error. The main benefits of MLIRT modeling are related to the possibility of estimating the quantity of the latent trait (e.g., motivation, satisfaction, achievement) at each level of analysis, conditional upon its quantity at other levels (e.g., classes or schools) of the data structure, and accounting for measurement error in making comparisons across groups by considering relevant predictors (Pastor, 2003). This means that the focus of the analysis can be on estimates of the latent trait at the individual level (students), group level (classes), or both. This requires that the proprieties of the latent traits have to be properly assessed at each level (e.g., student and class) and according to the aims of the analysis. Doing so ensures that there is construct validity across the levels and avoids spurious conclusions (see, for example, Forer & Zumbo, 2011; Linn, 2009; Zumbo & Forer, 2011). The peculiarities of multilevel IRT have a crucial role in studies whenever the main goal is to make comparisons across groups removing the effect of potential moderating factors or/and explanatory variables (Draper & Gittoes, 2004; Goldstein & Spiegelhalter, 1996) at different levels of analysis (students, schools, and teachers).

So, multilevel IRT modeling brings to fruition all of the advantages we gain from using IRT versus classical test theory (CTT; see De Ayala, 2013; Embretson & Reise, 2000; McEldoon, Cho, & Rittle-Johnson, 2012) and MLM versus single-level modeling (Hox, 2010; Peugh, 2014; Raudenbush & Bryk, 2002). Similar to Fox and Glas (2003), we use MLIRT to represent the strengths of this combination.

The MLIRT approach has a further advantage of providing more accurate estimates of the relationship between latent traits and predictors at different levels of analysis relative to a two-step MLM approach, that is (a) create raw or IRT scores on an outcome and (b) analyze multilevel data using scores from Step 1 as the outcome variable in subsequent analyses by simultaneously relying on the estimates of latent traits and amount of measurement error, item parameters, and covariates effects (Adams, Wilson, & Wu, 1997; Fox, 2007; Kamata, 2001; Pastor, 2003). This is especially true when the number of observed item responses for a respondent is low (i.e., short forms), and/or respondents are differing in numbers of items completed (Fox, 2007). An MLIRT modeling framework can improve measurement error estimates that get lumped into total raw scores or IRT scores that are generated using a two-step MLM approach (see, for example, Fox, 2007; Fox & Glas, 2003).

The goal of this article is to illustrate steps involved in conducting, interpreting, and presenting results from an application of MLIRT modeling without and with covariates using limited statistical jargon. Specifically, this article focuses on (a) specifying the research purpose and questions under investigation, (b) providing a brief overview of IRT, (c) assessing if MLM is necessary, (d) describing MLIRT for measurement description, and (e) demonstrating MLIRT for explanation (adding relevant covariates). All examples are discussed around one sample data set and instrument. The results within this article focus on estimates from the Stata 13 program Generalized Linear Latent and Mixed Models (GLLAMM; Rabe-Hesketh, Skrondal, and Pickles (2004b), but equivalent Stata 13 program structural equation modeling (SEM; StataCorp, 2013), M*plus* 7.4 (L. K. Muthén & Muthén, 1998-2015), and flexMIRT 3.0 (Cai, 2015) command and output files are provided as supplementary online materials at http://jea.sagepub.com/supplemental.

An Illustrative Example Using Sample Data From the Italian National Institute for the Evaluation of School System (INVALSI) Survey

To illustrate the MLIRT models without and with covariates, we considered a generated sample that reproduces the data structure and relationships among variables from an annual survey conducted by the INVALSI. The survey assesses skills and knowledge of students in the fifth year of the primary school (about age 10) enrolled in both private and public schools. For our purposes, we sampled one class from each school. The example data focuses on a multi-item (I = 25; coded as m1-m25) test that measures students' achievement in mathematics, wherein items are dichotomously scored (1 = correct, 0 = incorrect) from early adolescent students (J = 1,141; coded as id) nested within classes (K = 60; coded as class). Students did not move across teachers/classes and were situated in the same class/teacher. The average class size was about 19 students. In the following i (i = 1,…, I) denotes items, j (j = 1,…, J) indicates the students, and k (k = 1,…, K) represents the class to which they belong.

Specifying the Research Purpose and Questions Under Investigation

The MLIRT example analysis used herein examines response variation at both student and class levels and the influence of student gender (coded as gender; 0 = male, 1 = female) and average parents education (avg_ped) math achievement (coded as math). Similar to the approach used in Kamata (2001), Pastor (2003), and Peugh (2014), a model-building procedure will be used to answer the following research questions:

> **Research Question 1:** Are differences in mathematics achievement between classes relevant?
> **Research Question 2:** Are differences in the latent trait (math) between students explained by students' gender and parent average educational level (avg_ped)?

Brief Overview of IRT

Why use IRT? IRT is a probabilistic framework for the development and analysis of a multi-item instrument that has the potential to jointly shed light on items and person characteristics by conjointly linking item parameters and latent trait values on the same measurement scale. IRT models can be used for several aims, specifically to (a) choose items which effectively discriminate across individuals with a different magnitude of the latent trait; (b) select informative items to build up a briefer instrument and remove redundant items; (c) quantify the informative power of each item and an instrument to assess which segments of the latent trait are poorly measured; (d) measure individuals and item/item-categories (or item-categories for polytomous item response) location along the latent trait; (e) place individuals and item (or item-categories for polytomous data) characteristics on the same scale; (f) assess the minimum level of individual latent trait needed to likely provide a positive answer to an item or endorse a specific category of response to a certain item; (g) provide a measure of reliability that varies along the latent trait. Although a full description of IRT models is beyond the scope of this article and a listing of all the papers that have discussed this topic is impossible, interested learners are referred to De Ayala (2009), Edelen and Reeve (2007), Lord (1980), and Toland (2014) for further information.

Basic concepts and common dichotomous IRT models. Even though there are a variety of IRT models, for simplicity this study focuses on two IRT models for dichotomously scored item responses. The two most well-known IRT models for dichotomous item responses are the one-parameter logistic (1PL; Rasch, 1960) model and the two-parameter logistic (2PL; Birnbaum, 1968) model. The models are so called because of the number of item parameters each model contains. The 2PL model is the more general model and can be described by the following mathematical expression (Lord, 1980):

$$P\left(y_{ji}=1\,|\,\theta_j,b_i,a_i\right)=\frac{e^{a_i\left(\theta_j-b_i\right)}}{1+e^{a_i\left(\theta_j-b_i\right)}}, \tag{1}$$

where P is the probability that a randomly chosen person j with latent trait value θ will answer item i correctly or endorse it. The values of b_i and a_i are parameters characterizing item i, and e is the mathematical constant 2.71828. The relationship between a correct item response and the latent trait can be modeled using a logistic (S-shaped) function known as an item response function (IRF). This function specifies that as the level of the latent trait (math) increases, the probability of a correct response on an item will increase nonlinearly.

The two item parameters, a_i and b_i, are called item "discrimination" (a) parameter and item "location" or "difficulty" (b) parameter. The discrimination parameter (a_i) provides information on the steepness of the logistic function signaling the power of an item to differentiate across individuals with different levels of the latent trait. The item location parameter b_i signals the position of the item alongside the latent trait; specifically the minimum

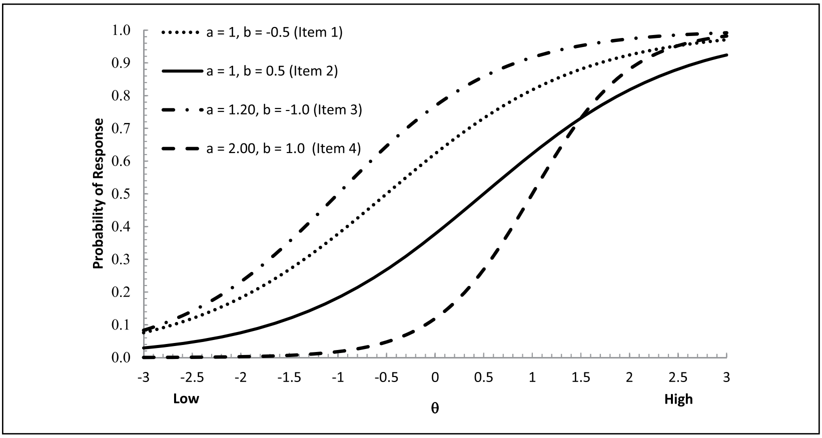

Figure 1. Example one-parameter (Rasch) and two-parameter logistic models item response functions for four hypothetical items dichotomously scored.
Note. a = discrimination; *b* = location. The horizontal axis represents the level of the latent trait (which has a standard normal distribution by construction), and vertical axis measures the probability of choosing or endorsing the category exactly true at a specified latent trait level.

amount of latent trait required to have a chance of 0.50 to answer it correctly. So, greater values of this parameter indicate that higher values of the latent trait are required to provide a positive answer to the item.

The values of the person parameter (θ), which identifies the individual position along the latent trait, and the item parameters are measured on the same metric. This links both parameters on the same continuum. If items have the same estimated discrimination power, then the parameter a_i becomes constant parameter (a) and the 2PL model reduces to the 1PL model. If the a parameter is set or fixed at 1 (not estimated), then the 1PL model is known as the Rasch model (Rasch, 1960).

Figure 1 plots four hypothetical IRFs. Items 1 and 2 depict items with equal discrimination as expected by a 1PL model, whereas Items 2 and 3 depict items with varying discrimination in accordance with a 2PL model. Notice that as one moves from low to high along the θ continuum, the probability to respond positively to an item increases, but is dependent on person location (θ)

Statistical frameworks. The basic IRT model can be viewed from two different perspectives: namely as a multivariate regression model (e.g., multiple items can be viewed as a set of outcome variables) or as a multilevel regression model (e.g., the set of items can be viewed as repeated measurements on the same student). In the multivariate regression framework, the basic IRT models can be set as a single-factor measurement model with categorical responses (StataCorp, 2013). The relationships between the observed responses to each item and the latent factor are described by I logistic regression equations, each of which is a function of an item intercept and a loading. Thus, IRT models can be specified with software which fits confirmatory factor analysis with categorical items such as M*plus* or the package for generalized structural equation models such as the Stata program SEM. Within this tutorial, we follow the multilevel regression approach, which is also the approach adopted in the Stata program GLLAMM. In both frameworks, IRT item parameters (a_i and b_i) are usually specified as fixed effects, whereas the person parameter θ, which is based on responses provided by the same person, can be specified as a fixed or random term, which is usually assumed to follow a normal distribution, $N\left(0,\sigma^2\right)$. Readers interested in replicating the analysis within a multivariate regression model or SEM framework are invited to read the supplementary materials containing data and command files that use the Stata program SEM, M*plus*, and flexMIRT.

Fixed- or random effects approach. The nesting of units in clusters (e.g., the same person responds to several items, or students belonging to the same class) is usually modeled by introducing fixed or random effects among the predictors of a regression model. In a fixed-effects approach, cluster affiliation is considered as a categorical covariate and a number of dummy variables, equal to the number of clusters minus 1 (i.e., a value of 1 is used for those units belonging to the same cluster on a given dummy variable and 0 elsewhere), are introduced as predictors. The effect of belonging to a cluster on the expected response is summarized by the related parameter estimate. As the number of clusters increases, the number of parameters that have to be estimated increases and the fixed-effects approach becomes inefficient. The random effects approach (more commonly known as a Multilevel approach) overcomes this issue by treating cluster effects as random terms (usually it is assumed that they are generated from a normal distribution) which assume the same values for units in the same cluster. The degree of similarity in responses of individuals in the same cluster (which are not explained by individual covariates) is considered by adding to the other predictors of the regression a random intercept which is shared by units belonging to the same cluster (e.g., responses of the same student or students in the same class). In a random effects framework, likely sources of dependence within higher level units can be easily taken into account by adding further random terms at higher levels of the data structure (e.g., classes in the same school). In this way, units belonging to the same cluster are treated as more alike than those which do not and therefore students in the same class and in the same schools share two common random terms. Moreover, a random effects formulation of the latent trait person parameters is recommended when a brief instrument is under analysis, when the number of individuals increases, or when the analysis focuses on person-level covariates. Ultimately, we adopt a random effects approach to specify the person parameter and to account for the nesting of items within students and students within classes.

Considering items as fixed parameters and adopting a random effects formulation of the person parameter, the most common IRT models for dichotomous and polytomous items can be classified as generalized linear or nonlinear mixed effects models (De Boeck & Wilson 2004; Fisher & Molenaar, 1995). This approach is fruitful as it can be used to extend IRT models in order to deal with a hierarchical or nested data structure, a latent variable that has a multidimensional structure (multidimensional IRT), and with covariates at person level (De Boeck & Wilson, 2004; Fisher & Molenaar, 1995; Pastor, 2003; Rijmen & Briggs, 2004).

Link between IRT and multilevel logistic model. At the turn of the 21st century, a number of pioneering papers (see, for example, Kamata, 2001; Singer, 1998) show that the class of Rasch models (i.e., where a is set at 1 and not estimated) for dichotomous and polytomous item response data can be parameterized as two-level logistic regression models. In this framework, a random intercept is specified at the person level (Level 2) and item indicator (dummy) variables are specified as predictors at the item response level (Level 1).

In order to adopt such an approach to estimate IRT parameters, the data have to be converted from wide (multivariate, person-level) format (see Appendix Table A1) to long (stacked, person-period) format (see Appendix Table A2) wherein the responses to the 25 items (m1-m25) of the 1,141 individuals are stacked in a unique variable m and each individual has a number of records/lines/rows equal to the number of items (see INVALSIMATHLONG.csv). The variable m (which has length or rows = 28,525 = 1,141 by 25 and assumes values 1 or 0) is used as the dependent variable in a multilevel logistic regression model containing as predictors 25 binary variables (x_i, for i in 1,…, 25; see Appendix Table A3). Each binary variable has a value 1 when the response in the m variable is associated to item M$_i$, otherwise 0. When the model is estimated, the related coefficient parameters of the $\mathbf{x}_i$ variables are item intercepts.

In the formulation of the IRT model as a two-level multilevel logistic regression model (Adams et al., 1997; Kamata, 2001), the person parameter θ_j is treated as a random term, which is made up of responses to items provided by the same person. This term captures the unexplained (residual) variability in item responses attributable to differences in person latent trait after controlling for item characteristics (i.e., item location parameters).

The Rasch model can be specified as a two-level multilevel logistic regression model with item responses (Level 1 units) nested within students (Level 2 units). In this model, item indicators are covariates, which are the x_i binary variables that identify responses to the same item and expressed as follows (Jeon & Rabe-Hesketh, 2012; Kamata, 2001):

$$P\left(m_{ij} = 1 \mid \theta_j, \alpha_i\right) = \frac{e^{\alpha_i x_i + \theta_j}}{1 + e^{\alpha_i x_i + \theta_j}}. \tag{2}$$

In Equation 2, the probability that student j provides a correct response ($m_{ij} = 1$) to item i depends on an item intercept $\alpha_{i,}$ and a common latent trait (random term), which takes the same value for all the responses related to the same person θ_j. Equation 2 can be obtained from Equation 1 by setting $-\alpha_i = \mathrm{b}_i$ and $a = 1$. The item intercept α_i can be interpreted as an easiness parameter, the larger its value the greater the probability to get a correct response. The item loading λ can be interpreted as a discrimination parameter, the larger the value the greater the relationship between the item and latent trait. Note, item loading is not estimated in this formulation as the value is fixed at 1 as is done in Rasch modeling. The model in Equation 2 can be estimated with any software which allows users to specify a generalized linear mixed effects model, such as Stata, M*plus*, MLwiN, flexMIRT, SPSS, and SAS procedures PROC NLMIXED or PROC GLIMMIX.

The 2PL model in Equation 1 can be expressed in terms of an item intercept α_i and loading λ_i unique for each item as (Skrondal & Rabe-Hesketh, 2004)

$$P\left(m_{ij} = 1 \mid \theta_j, \alpha_i, \lambda_i\right) = \frac{e^{\alpha_i x_i + \lambda_i \theta_j}}{1 + e^{\alpha_i x_i + \lambda_i \theta_j}}, \tag{3}$$

where $\alpha_i = -a_i b_i$ and $\lambda_i = a_i$ (McDonald, 1999). The item intercept α_i can be expressed in terms of "location/difficulty parameter" as $b_i = -\alpha_i / \lambda_i$.[1] As the discrimination-location form does not generalize well to the truly multilevel IRT model, we adopted the intercept-loading parameterization for this model and all remaining IRT models. Equation 3 is an IRT model with a unidimensional latent trait expressed in intercept-slope parametrization. The 2PL model is also a generalized nonlinear model since in Equation 3 the loading parameter is multiplied by the person parameter (Jeon & Rijmen, 2015).

To identify Equation 3, it is required to fix the variance of the random term θ_j equal to 1 or a loading equal to 1. If we suppose that all items have the same discrimination power of 1, then Equation 3 simplifies to Equation 2.

The 2PL model can be expressed in the generalized nonlinear mixed model framework and estimated using routines for mixed models that allow varying loadings, such as Stata programs GLLAMM and SEM, the PROC NLMIXED procedure in SAS as illustrated by Jeon and Rabe-Hesketh (2012), Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003), and StataCorp (2013). Another option is to use software which relies on Bayesian methods such as WinBUGS (Cho & Suh, 2012).

Assessing If an MLM is Necessary

When we observe a data structure with nested units, it is necessary to use MLM, but not a sufficient condition. Many researchers (Raudenbush & Bryk, 2002; Reise, Ventura, Nuechterlein, & Kim, 2005) consider the estimation of an unconditional MLM to allow the computation of variance partition coefficient (VPCs) or intraclass correlation coefficients (ICCs) for an outcome variable as an essential step prior to any type of MLM. A multilevel analysis is required only if there is a relevant amount of variation in the outcome variable explained by higher level units (e.g., classes). Although no definitive rule can be made about what defines the minimum relevant amount of variance needed to justify an MLM, research on MLM suggest at minimum VPCs or ICCs around 5% (see Glaser & Hastings, 2011; Hayes, 2006; p. 394; Kreft & de Leeuw, 1998, pp. 9-10). It is important to mention that just because data are nested does not mean a multilevel analysis is needed unless it fits the research questions at hand for explaining the outcome variable variance. The main issue with ignoring a nested data structure is the bias in standard error estimates in the statistical tests and confidence intervals. So, the bias can be corrected either with an MLM approach or using sampling weights. In the following, an exploratory MLM analysis was carried out to shed light on the relevance of class-level variability in the INVALSI data with respect to responses provided in items addressed to measure students' achievement in mathematics.

To address the first research question, we initially assessed the intensity of intraclass variability across the responses to each item provided by students nested in the same class. This means that the data only consists of two levels (students within class). This was done by introducing a random intercept at the class level in an unconditional two-level MLM and assessing the size of the variance partitioning coefficient (VPC). In this way, the variability in students' responses in an item is split into two parts: the amount attributable to differences across students within classes and the amount due to differences across classes. A two-level multilevel logistic model was used to link (relate) the dichotomous responses for single item to the overall item intercept (γ_{oo}), plus a random error intercept ($\mu_{0k} \sim N[0, \sigma_{\mu}^{2}]$) expressed as (Hox, 2010)

$$P\left(m_{jk}=1\right)=\frac{e^{\gamma_{oo}+\mu_{0k}}}{1+e^{\gamma_{oo}+\mu_{0k}}}. \tag{4}$$

Equation 4 quantifies the probability of a correct response for person *j* (Level 1) in class *k* (Level 2) on a given item. Using the variance of the random intercept (σ_{μ}^{2}), we can estimate the amount of between-classes variability. In a logistic regression model, the variance of the Level 1 unit is fixed to the variance of the logistic distribution, $\pi^2/3$ (Steele, 2009).[2] In our case, the VPC is the ratio of between-class variance to total variance (i.e., between-class variance + within-class variance) in item *i* ($\text{VPC}_i = \sigma_{\mu}^{2} / [\sigma_{\mu}^{2} + \pi^2/3]$; Snijders & Bosker, 2012). VPC values close to 0 indicate that variation is almost all due to differences among students than among classes which is akin to meeting the assumption of sampling at the individual level compared with sampling at the cluster level. In comparison, values close to 1 indicate that within-group differences in students' responses in the same class are nearly null and due to differences between classes. In order to estimate the VPC for each item, data were presented in a wide format (see Appendix Table A1).

A common way to test the significance of the between-class variability or VPC is to use a likelihood ratio test (LRT), which is the difference in the −2 log likelihood (−2LL) of two nested models. In our case, this would be the reduced model (single-level logistic regression model) minus the full model (two-level multilevel logistic regression model). The LRT is χ^2 distributed with degrees of freedom equal to the difference in the number of parameters estimated between the two nested models. We tested the significance of the LRT to assess if the multilevel analysis was worthwhile. Importantly, though, the LRT is conservative for comparing models, thus a common way to proceed is to half the associated *p* values (Jak, Oort, & Dolan, 2014; Snijders & Bosker, 2012).

Researchers have also suggested that a design effect (i.e., design effect = 1 + (average cluster size – 1) × VPC) larger than 2 necessitates a multilevel analysis (B. O. Muthén & Satorra, 1995). This means that for our data with an average class size of 19.017, the VPC should be larger than .056 or 5.6%. Table 1 provides for each item (m1-m25) the LL for each model along with the corresponding LRT, between-class variance, VPC, and design effect.

Results show that the values of the VPC ranged between 10% and 49% with an average of 22.1%. All items show values of the VPC are significantly different than zero and greater than 5.6%. The values of the design effect (always greater 2) in the last column of Table 1 indicate that a substantial (nontrivial) amount of each item's variation is attributable to differences among the class and are not explained by student characteristics alone. As such, an MLM approach is appropriate for understanding the influence of belonging to a class in explaining differences in mathematics achievement

Table 1. Log Likelihood for Single-Level Logistic Regression Model (Reduced Model) and Two-Level Multilevel Logistic Regression Model (Full Model) for Each Item Along With Likelihood Ratio Test, Between-Class Variance (σ_μ^2), Variance Partitioning Coefficient, and Design Effect.

Item	Reduced model LL	Full model LL	LRT	σ_μ^2	VPC	Design effect
m1	−512.21	−495.01	34.40	0.66	.17	4.00
m2	−749.61	−732.18	34.86	0.41	.11	2.98
m3	−789.55	−754.47	70.17	0.63	.16	3.88
m4	−535.76	−524.41	22.71	0.46	.12	3.20
m5	−783.11	−751.72	62.77	0.60	.15	3.78
m6	−658.35	−629.14	58.41	0.73	.18	4.28
m7	−766.08	−697.68	136.80	1.13	.26	5.60
m8	−789.91	−738.53	102.75	0.90	.21	4.85
m9	−702.57	−660.42	84.30	0.84	.20	4.65
m10	−507.33	−497.61	19.43	0.48	.13	3.28
m11	−782.15	−762.30	39.69	0.41	.11	2.98
m12	−788.14	−723.15	129.98	1.14	.26	5.63
m13	−790.81	−693.19	195.23	1.68	.34	7.08
m14	−780.05	−718.00	124.09	1.08	.25	5.45
m15	−788.41	−734.06	108.71	0.96	.23	5.06
m16	−784.87	−753.98	61.77	0.58	.15	3.69
m17	−745.11	−676.70	136.82	1.11	.25	5.54
m18	−786.23	−626.13	320.18	3.17	.49	9.83
m19	−723.93	−703.77	40.33	0.48	.13	3.29
m20	−787.40	−744.22	86.35	0.77	.19	4.41
m21	−726.72	−655.41	142.63	1.35	.29	6.23
m22	−788.41	−742.68	91.47	0.75	.19	4.35
m23	−618.40	−597.58	41.65	0.62	.16	3.84
m24	−430.62	−425.74	9.76	0.36	.10	2.78
m25	−621.96	−609.85	24.23	0.40	.11	2.94

Note. VPC = σ_μ^2 /($(\sigma_\mu^2$ + $\pi^2/3)$. $\pi^2/3$ = 3.29 or variance of logistic distribution. Design effect = 1 + (Average cluster size − 1) × VPC; Average class size = 19.017. Average design effect across items = 4.603. All LRTs are significant at α = .001. LRT = likelihood ratio test; VPC = variance partitioning coefficient; LL = log likelihood.

between students and provides the information needed for answering Research Question 1.

Alternatively we jointly consider the responses provided by students to the 25 items and the clustering of students within classes in order to assess the average amount of variability attributed to class differences in the responses to the mathematics test. Within an MLM framework, we can do this by fitting a three-level multilevel logistic regression model, as is done in the Stata program GLLAMM: item responses (Level 1 units, 1,141 by 25 = 28,525) within students (Level 2 units, 1,141) across classes (Level 3 units, 60). To carry on with the three-level analysis in Stata program GLLAMM, the data set has to first be set in long format (see Appendix Table A2).

Putting the data set in long format allows us to simultaneously estimate the average shared variability in the entire sample occurring at person (Level 2) and class level (Level 3) by fitting a (null) three-level multilevel logistic regression model. This model links the dichotomous responses to item *i* (Level 1) of person *j* (Level 2) in class *k* (Level 3) to an overall intercept (γ_{oo}), plus two random intercepts, one at the student level $(\theta_{jk} \sim N[0,\ \sigma^2_{\theta^{(s)}}])$ and one at the class level, $\left(\theta_{0k} \sim N\left[0, \sigma^2_{\theta^{(c)}}\right]\right)$ expressed as (Skrondal & Rabe-Hesketh, 2004)

$$P\left(m_{ijk} = 1\right) = \frac{e^{\gamma_{oo} + \theta_{jk} + \theta_{0k}}}{1 + e^{\gamma_{oo} + \theta_{jk} + \theta_{0k}}} \tag{5}$$

Equation 5 quantifies the probability of a correct response to item *i* for person *j* in class *k*. Equation 5 is different from Equation 4 in that the random intercept at the class level can be interpreted in terms of performance in class achievement due to class.

In Equation 5, the overall variability in the responses (variable *m* in Appendix Table A2) is partitioned into three components: the between-items within-individuals component (at Level 1) that is fixed at $\pi^2/3$ (see Note 2), the between-students' within-class component $\sigma^2_{\left(\theta^{(s)}\right)}$ (at Level 2), and the between-class component $\sigma^2_{\left(\theta^{(c)}\right)}$ (at Level 3). We refer to the model expressed in Equation 5 as the three-level variance component (VC) model.

Using an LRT, the significance of the class-level component $\sigma^2_{\left(\theta^{(c)}\right)}$ can be tested by comparing two nested VC models: the model that considers the nesting of students in classes (three-level VC model) versus the model which ignores the clustering of students within classes (two-level VC model).

Table 2 provides a comparison between the two-level and three-level VC models for response variable *m* (Data Set in Long Form) and the related VPCs and design effects. Results suggest that the overall variability is split into two parts: approximately 11.5% (.390 / [.390 + .477] × 100) due to the class component (Level 3 variability) and 9.4% (.477 / [.390 + .477] × 100) due to the student component (Level 2 variability). Similarly, the design effect is split into two parts, respectively, 3.28 at the student level and 3.072 at the class level. The LRT also shows that including the class level improves

Table 2. Comparisons Between Two-Level and Three-Level VC Models for Response Variable *m* (Data Set in Long Form).

VC model	LL	LRT	$\sigma^2_{\theta^s}$	$\sigma^2_{\theta^c}$	VPC$^{(s)}$	VPC$^{(c)}$	Design effect$^{(s)}$	Design effect$^{(c)}$
Two level	−18,178.68		.836		.202		5.848	
Three level	−17,854.91	647.55***	.390	.477	.094	.115	3.280	3.072

Note. VC = variance component; LL = log likelihood; LRT = likelihood ratio test; $\sigma^2_{\theta^s}$ = between-students' variance component; $\sigma^2_{\theta^c}$ = between-class variance component; VPC = variance partitioning coefficient; $\text{VPC}^{(s)} = \sigma^2_{\theta^s} / \left(\sigma^2_{\theta^s} + \sigma^2_{\theta^c} + \pi^2 / 3\right)$. $\text{VPC}^{(c)} = \sigma^2_{\theta^s} / \left(\sigma^2_{\theta^s} + \sigma^2_{\theta^c} + \pi^2 / 3\right)$; $\pi^2/3 = 3.29$ or variance of logistic distribution; Design effect$^{(s)}$ = 1 + (25 – 1) × VPC$^{(s)}$; Design effect$^{(c)}$ = 1 + (19.017 – 1) × VPC$^{(c)}$. LRT is significant at $\alpha = .001$.

the explanation of the variability in item responses versus a two-level model that ignores the class component. If we ignore the class component, there is an increase in the variability of the random intercept at the student level (Level 2 variability, $\sigma^2_{\left(\theta^{(s)}\right)}$), and this variability is wrongly attributed to individual differences. In conclusion, a three-level model should be used to appropriately explain divergences in mathematics achievement across students.

MLIRT for Measurement Description and Explanation

Why use MLIRT? Within the MLIRT framework, we gain all the potential of IRT along with four additional features of being able to (a) assess how much an item discriminates between classes and within individuals; (b) assess how much of the variability in the responses is due to individuals and to classes; (c) estimate a value of the latent trait at class level; (d) use the estimates of the latent trait at class level as indicators (value-added measures) to measure effectiveness removing the effects of the heterogeneity in students' characteristics (Goldstein & Spiegelhalter, 1996); (e) assess the amount of information (reliability) of the latent trait at both class and student levels.

Dichotomous MLIRT models. By extending the 2PL model in Equation 3, we can consider the nesting of students in the *K* classes to obtain a multilevel logistic regression model and call it a multilevel 2PL (MLM-2PL) model. Now, the MLM-2PL model can be used to express the probability to respond correctly to item *i* as

$$P\left(m_{ijk}=1\mid\theta,\alpha,\lambda\right)=\frac{e^{\alpha_i x_i+\lambda_i^{(s)}\theta_{jk}+\lambda_i^{(c)}\theta_{.k}}}{1+e^{\alpha_i x_i+\lambda_i^{(s)}\theta_{jk}+\lambda_i^{(c)}\theta_{.k}}},\tag{6}$$

where $\boldsymbol{\theta}$ represents the vector of the latent trait values at the student level ($\theta_{jk}\sim N[0,\ \sigma^2_{\theta^{(s)}}]$) and at the class level, $\left(\theta_{0k}\sim N\left[0,\sigma^2_{\theta^{(c)}}\right]\right)$. Also, $\boldsymbol{\lambda}$ is the vector containing the loadings at the student level ($\lambda_i^{(s)}$) and class level ($\lambda_i^{(c)}$) and as before α_i is the item intercept. The student-level loading measures item *i*'s capability to differentiate across students (between-students within-classes) with different achievement in math conditional upon the class level of the latent trait, whereas the loading at the class level provides "value-added" information on item *i*'s capacity to discriminate between classes on the basis of students' achievement in math. It is worth highlighting that in MLM, the sample size of units at each level of analysis (e.g., item responses [Level 1], students [Level 2], and classes [Level 3]) is relevant in order to assess the power of statistical tests related to fixed and random parameters. Moreover, the appropriate sample size at higher levels (Level 2 and Level 3) depends on the parameter(s) being tested (e.g., fixed vs. random) and the main focus of the research question(s) (Snijders, 2005). If the aim of an analysis is to test fixed effects, then the sample size of higher level units is not relevant. A guide and statistical software program for determining statistical power and sample size in MLM is provided by Browne, Golalizadeh Lahi, and Parker (2009). However, to our knowledge, there is not a program readily available for directly testing statistical power in the MLIRT context.

To identify the MLM-2PL model, a loading at each level of analysis (students and classes) has to be constrained to 1 and then variances of the related random terms are freely estimated. Or, the model can be identified by constraining the variances of the random terms to 1. In our application of the MLIRT models for comparing the relative contribution of each level of analysis to the overall variability and when comparing competing models we allowed the variances of the random terms to be freely estimated. When conducting MLIRT solely for measurement description purposes, the variances of the random terms were fixed to 1 to standardize the metric and improve interpretability.

If loadings $\lambda_i^{(c)}$ at class level are supposed to have the same capacity to discriminate across classes, then all loadings are set equal to 1 at the class level and the MLM-2PL model reduces to the multilevel 2PL constrained (MLM-2PLC) model which can be expressed as

$$P\left(m_{ijk} = 1 \mid \theta, \alpha, \lambda_i^{(s)}\right) = \frac{e^{\alpha_i x_i + \lambda_i^{(s)} \theta_{jk} + \theta_{.k}}}{1 + e^{\alpha_i x_i + \lambda_i^{(s)} \theta_{jk} + \theta_{.k}}}. \tag{7}$$

To test if the MLM-2PLC model works, a comparison needs to be made with the model described in Equation 6 using a LRT. If the specification of loadings at the class level do not improve the model fit, then it means that the loadings at the class level do not differ in a significant way from 1 and the loadings are the same for the *i* items. The model expressed in Equation 7 may possibly be realistic when classes perform differently at the instrument level, but not at the item level. Finally, if items have the same power to discriminate across students, then loadings at the student level are constant across items and equal to 1 (i.e., $\lambda_i^{(s)} = \lambda^{(s)} = 1$). Thus, Equation 7 reduces to the multilevel Rasch (MLM-Rasch) model expressed as

$$P\left(m_{ijk} = 1 \mid \theta, \alpha\right) = \frac{e^{\alpha_i x_i + \theta_{jk} + \theta_{.k}}}{1 + e^{\alpha_i x_i + \theta_{jk} + \theta_{.k}}}. \tag{8}$$

Equation 8 represents the probability of answering correctly an item as a function of two random terms, namely θ_{jk} and θ_k, which respectively capture the difference between the *k* class's level of the latent trait and the overall mean, and the difference in math achievement of student *j* in class *k* with respect to the overall class *k* achievement. Classes with higher values of θ_k tend to have students with higher math achievement, whereas lower values of θ_k signal that students have on average lower levels of math achievement. The item intercept $(-\alpha_i = b_i)$ signals the difference in the probability of a correct answer due to a difference in item difficulty.

Notice that in the MLM-2PLC model and MLM-Rasch model that in order to assess the relative contribution of both levels to the overall variability, the variance of the random terms, which capture students and class variability, are estimated and the values of the loading parameter are fixed equal to 1. Note, Equation 5 differs from Equation 8 in that it has an overall intercept across all items, whereas Equation 8 allows each item to have its own unique intercept.

The MLM-2PL model in Equation 6 is estimated in the Stata program GLLAMM as a three-level model with loadings at student (Level 2) and class level (Level 3) which represent the within-class between-students loadings and the between-classes loadings (see Rabe-Hesketh et al., 2004b; Skrondal & Rabe-Hesketh, 2004).

Similar to traditional dichotomous IRT models, dichotomous MLIRT models make similar assumptions. One assumption is that the IRT model is modeling the correct number of latent traits that exist in the data at both student- and class level. A second assumption is conditional independence. This implies that conditional upon the value of the latent trait at student- and class-level responses to the item are due only to the latent trait(s) being modeled and nothing else. A third assumption is that the model appropriately represents the functional form of the data. Multilevel dimensionality can be assessed using exploratory and confirmatory factor analytic methods via Stata and M*plus*. Unfortunately, direct tests of the second and third assumptions are not standard features available in Stata and M*plus* (see Jiao, Kamata, Wang, & Jin, 2012; Ravand, 2015). Nonetheless, testing these assumptions is beyond the scope of this article since it would add unnecessary length. It is suggested that an extension of Yen's (1984) Q3 index be used to examine the conditional independence assumption. Q3 is the correlation between the residuals for a pair of items where the residual is the difference between observed response (0/1) and the predicted response probability from the MLIRT model. So, the residual for student *j* in class *k* on item *i* is as follows: $d_{jik} = y_{jik} - p_{ijk}(\theta_{ik}\theta_k)$, whereas for the same person the residual for item *m* is $d_{jmk} = y_{jmk} - p_{jmk}(\theta_{ik}\theta_k)$. The closer Q3 is to zero the stronger the evidence that the assumption is tenable. Yen suggested using a cut value of 0.2. The standard has remained in place for more than 20 years, and although not perfect, continues to be shown as the optimal criterion for assessing this assumption with Yen's Q3 index.

MLIRT results. Table 3 provides a summary of the characteristics of the five IRT models we focus on in this section. We investigated the advantages of fitting an MLIRT model in the analysis of students' achievement in math by answering the following question: Are differences in mathematics achievement between classes relevant? In order to answer this question, several MLIRT models (MLM-Rasch, MLM-2PLC, and MLM-2PL models) were fit to each item on the mathematics achievement test and results compared with models that have similar proprieties, but ignore the multilevel data structure (Rasch, 2PL). Table 4 summarizes the item intercept and factor loading parameters for each of the five models and the model statistics. The models were identified by setting the first item loading to 1.

Table 4 shows that across the models there are slight differences in the estimates of the item intercept. Also, the five models provide similar rankings of the items in terms of item intercept as the Spearman correlation coefficient estimates (r_s) between pairs of rankings of the estimates of intercept parameters obtained with different models ranged between .99 (2PL vs. MLM-2PL) to 1.00 (Rasch vs. ML Rasch). Recall, the relationship between item intercepts and location parameters can be used to interpret the item intercept (α)

Table 3. Characteristics of the Five Descriptive IRT Models.

Model characteristics	Rasch	2PL	MLM-Rasch	MLM-2PLC	MLM-2PL
Item parameter					
Intercept	Yes	Yes	Yes	Yes	Yes
Factor loading					
@ student level		Yes		Yes	Yes
@ class level					Yes
Random effect					
@ student level	Yes	Yes	Yes	Yes	Yes
@ class level			Yes	Yes	Yes

Note. IRT = item response theory; 2PL = two-parameter logistic; MLM = multilevel modeling; MLM-2PLC = multilevel 2PL constrained; MLM-2PL = multilevel 2PL.

results in Table 4 in terms of "item difficulty parameters" rather than "item intercept" by multiplying each item intercept by −1.

The specification or not of the multilevel structure does not seem to have a relevant impact on the estimates of the item intercept parameters. However, the drawback of omitting the multilevel structure in the IRT specification is greater with respect to the estimates of the item loading or discrimination parameters as the Spearman correlation estimates ranged between .75 and .85 between the 2PL and MLM-2PL models. The specification of between-classes factor loadings in the MLM-2PL (class level or Level 3 in GLLAMM) does not have a relevant impact on the within individuals loadings ranking (student level or Level 2 in GLAMM) for models that account for the multilevel data structure. For instance, the Spearman correlation coefficient estimate for item loadings estimates in MLM-2PLC and MLM-2PL was .97.

Now, focusing on the variances of the random terms at class- and student level, the main findings can be summarized as follows:

1. When the nesting of students in classes is ignored, the amount of variance due to differences in students' achievement is overestimated. The VPC(s) for the Rasch and 2PL models show that the between person parameter accounts for 26% and 31%, respectively, of the variability in the responses. However, the respective multilevel models, which consider the nesting of students in classes, have VPC(s) of 13% and 20%, respectively.
2. The variability due to individual differences is split into two parts when the class-level random term is introduced to take into account the nesting of students in classes: the within-class between-students'

Table 4. Summary of Item and Model Characteristics for Five IRT Model Solutions.

	Rasch		2PL				MLM-Rasch		MLM-2PLC				MLM-2PL					
Item	α_i	*SE*	α_i	*SE*	$\lambda_i^{(s)}$	*SE*	α_i	*SE*	α_i	*SE*	$\lambda_i^{(s)}$	*SE*	α_i	*SE*	$\lambda_i^{(s)}$	*SE*	$\lambda_i^{(c)}$	*SE*
1	1.95	0.09	2.10	0.13	1.00		1.93	0.13	2.06	0.15	1.00		2.13	0.181	1.00		1.00	
2	0.69	0.08	0.60	0.07	0.43	0.07	0.68	0.12	0.64	0.12	0.40	0.09	0.58	0.080	0.42	0.09	0.369	0.10
3	−0.10	0.07	−0.09	0.06	0.30	0.06	−0.12	0.12	−0.10	0.12	−0.03	0.07	−0.10	0.099	0.00	0.07	0.617	0.13
4	1.85	0.09	1.87	0.11	0.85	0.13	1.82	0.13	1.84	0.14	0.78	0.14	1.84	0.143	0.80	0.15	0.758	0.18
5	0.31	0.07	0.32	0.07	0.77	0.11	0.29	0.12	0.29	0.12	0.59	0.11	0.30	0.128	0.59	0.11	0.853	0.16
6	1.27	0.08	1.32	0.09	0.89	0.12	1.25	0.13	1.27	0.13	0.80	0.14	1.29	0.141	0.77	0.13	0.865	0.18
7	−0.50	0.07	−0.50	0.08	0.99	0.13	−0.52	0.12	−0.51	0.12	0.67	0.12	−0.52	0.152	0.74	0.13	1.053	0.19
8	−0.08	0.07	−0.05	0.08	1.08	0.14	−0.10	0.12	−0.11	0.12	1.03	0.16	−0.10	0.136	1.05	0.16	0.893	0.17
9	−1.00	0.08	−1.04	0.09	1.04	0.13	−1.02	0.13	−1.09	0.13	1.01	0.16	−1.08	0.132	0.99	0.16	0.790	0.15
10	1.97	0.09	1.86	0.10	0.65	0.11	1.95	0.13	1.86	0.13	0.50	0.12	1.87	0.145	0.50	0.12	0.812	0.19
11	0.33	0.07	0.33	0.07	0.76	0.10	0.31	0.12	0.31	0.12	0.68	0.12	0.30	0.112	0.66	0.12	0.701	0.14
12	−0.15	0.07	−0.13	0.08	0.91	0.12	−0.17	0.12	−0.18	0.12	0.95	0.15	−0.17	0.119	0.93	0.15	0.741	0.14
13	0.05	0.07	0.08	0.08	0.95	0.12	0.03	0.12	0.04	0.12	0.88	0.14	0.04	0.129	0.88	0.14	0.847	0.17
14	−0.32	0.07	−0.28	0.06	0.42	0.07	−0.34	0.12	−0.30	0.12	0.09	0.07	−0.30	0.108	0.12	0.07	0.698	0.14
15	−0.14	0.07	−0.12	0.08	1.19	0.15	−0.16	0.12	−0.17	0.12	1.05	0.16	−0.16	0.148	1.11	0.17	0.995	0.18
16	−0.24	0.07	−0.21	0.07	0.75	0.10	−0.25	0.12	−0.25	0.12	0.78	0.13	−0.25	0.102	0.73	0.12	0.596	0.12
17	−0.70	0.08	−0.71	0.08	1.00	0.13	−0.72	0.13	−0.72	0.12	0.74	0.13	−0.73	0.155	0.76	0.13	1.066	0.19
18	−0.21	0.07	−0.19	0.09	1.30	0.16	−0.22	0.12	−0.22	0.12	0.74	0.12	−0.06	0.286	0.91	0.15	2.145	0.40
19	0.89	0.08	0.93	0.08	0.90	0.12	0.87	0.13	0.95	0.13	1.09	0.17	0.90	0.106	1.03	0.16	0.523	0.12

(continued)

Table 4. (continued)

	Rasch		2PL				MLM-Rasch		MLM-2PLC				MLM-2PL					
Item	α_i	*SE*	α_i	*SE*	$\lambda_i^{(s)}$	*SE*	α_i	*SE*	α_i	*SE*	$\lambda_i^{(s)}$	*SE*	α_i	*SE*	$\lambda_i^{(s)}$	*SE*	$\lambda_i^{(c)}$	*SE*
20	0.21	0.07	0.24	0.08	0.91	0.12	0.20	0.12	0.21	0.12	0.87	0.14	0.21	0.123	0.85	0.14	0.785	0.15
21	0.87	0.08	0.91	0.08	0.90	0.12	0.85	0.13	0.85	0.12	0.73	0.13	0.91	0.154	0.73	0.13	1.036	0.21
22	0.18	0.07	0.20	0.07	0.81	0.11	0.17	0.12	0.18	0.12	0.98	0.16	0.16	0.095	0.97	0.16	0.483	0.11
23	1.47	0.08	1.76	0.12	1.24	0.17	1.45	0.13	1.91	0.17	1.73	0.27	1.86	0.164	1.67	0.26	0.675	0.16
24	2.32	0.10	2.69	0.17	1.21	0.18	2.29	0.14	2.85	0.21	1.57	0.26	2.79	0.205	1.54	0.25	0.641	0.18
25	1.45	0.08	1.36	0.09	0.64	0.10	1.43	0.13	1.37	0.13	0.53	0.11	1.35	0.123	0.51	0.11	0.715	0.16
$\sigma^2_{\theta^{(s)}}$	1.16	0.06	1.51	0.31			0.57	0.04	1.08	0.27			1.09	0.27				
$\sigma^2_{\theta^{(c)}}$							0.62	0.12	0.58	0.12			0.95	0.36				
#pars	26		50				27		51				75					
LL	−16,149.84		−16,008.65				−15,909.32		−15,750.19				−15,648.45					j
$VPC_{(s)}$	0.26		0.31				0.13		0.22				0.20					
$VPC_{(c)}$							0.14		0.12				0.18					

Note. IRT = item response theory; 2PL = two-parameter logistic; MLM = multilevel modeling; MLM-2PLC = multilevel 2PL constrained; MLM-2PL = multilevel 2PL; LL = log likelihood; VPC = variance partitioning coefficient.

variability and the between-classes' variability. For instance, for the MLM-2PL model, the VPC(s) and VPC(c) were 0.22 and 0.18, respectively.

3. Comparing the MLM-Rasch model with the MLM-2PL model, we can see that the introduction of factor loadings at the student level allows for a better representation of the within-students between-classes differences in achievement, whereas the introduction of the between-classes loadings allows to highlight divergences between classes. According to the results from the MLM-Rasch model (which does not consider the discrimination power of the items), about 13% of the variance is due to differences in students and about 14% to differences at class level, whereas for the MLM-2PLC model, the overall variability due to students and classes increases to 22% and 12%, respectively. Furthermore, introducing the between-classes loadings in the MLM-2PL model (which differentiates between the discrimination power of the items between and within classes) the VPC(c) goes up to 18%.

To find the best fitting IRT model, a series of LRTs were conducted and the results showed that the MLM-2PL model had the best fit to the data. For instance, the LRT for the comparison of the MLM-2PLC model to the MLM-2PL model was $-2(-15{,}750.19 - -15{,}648.45) = 203.48$ with $df = 75 - 51 = 24$ which has an associated p value $< .001$. Therefore, the LRTs show that including the class-level loadings improves the explanation of the variability in item responses. In conclusion, an MLM-2PL model is recommended to appropriately represent the data.

MLIRT IRF. Within the MLIRT framework, the probability of responding correctly to an item is a function of both the class parameter θ_k and the person parameter θ_{jk}. The class parameter quantifies the effect of belonging to class k on student math achievement, whereas the latter represents the effect of the student component on the observed math achievement score. To jointly represent the item response probabilities as a function of θ_k and θ_{jk}, a tridimensional plot can be used. However, such a plot can be difficult to interpret and is not easily created, especially when the number of nested levels increases and dimensionality increases. Furthermore, there currently is not a standard approach for graphically representing multilevel IRT functions (Vector Psychometric Group, 2015). Therefore, we believe it is helpful to grasp differences in function behavior in two dimensions. To do so we create IRFs at student- and class level conditioned at critical points at the class level or student level of the latent trait (math achievement) continuum; we call these

Table 5. Summary of MLM-2PL With Covariates.

Item	α_i	*SE*	$\lambda_i^{(s)}$	*SE*	$\lambda_i^{(c)}$	*SE*
1	2.20	0.19	1		1.00	
2	0.61	0.08	0.43	0.10	0.37	0.09
3	−0.10	0.10	0.00	0.07	0.59	0.12
4	1.89	0.15	0.83	0.15	0.72	0.17
5	0.35	0.13	0.63	0.12	0.82	0.16
6	1.35	0.14	0.86	0.15	0.78	0.17
7	−0.46	0.15	0.79	0.14	1.01	0.19
8	−0.02	0.14	1.10	0.18	0.85	0.16
9	−1.01	0.13	1.04	0.17	0.73	0.14
10	1.91	0.15	0.54	0.12	0.76	0.18
11	0.35	0.11	0.69	0.12	0.66	0.13
12	−0.11	0.12	0.96	0.15	0.72	0.14
13	0.11	0.13	1.02	0.17	0.74	0.15
14	−0.29	0.11	0.13	0.08	0.68	0.14
15	−0.08	0.15	1.15	0.18	0.94	0.18
16	−0.20	0.10	0.76	0.13	0.56	0.12
17	−0.67	0.16	0.77	0.13	1.02	0.19
18	0.02	0.28	0.97	0.17	2.03	0.38
19	0.98	0.11	1.12	0.18	0.46	0.11
20	0.27	0.13	0.88	0.14	0.76	0.15
21	0.96	0.15	0.78	0.14	0.96	0.20
22	0.24	0.10	1.07	0.17	0.42	0.10
23	1.97	0.18	1.75	0.28	0.61	0.15
24	2.86	0.21	1.58	0.26	0.58	0.17
25	1.39	0.12	0.57	0.12	0.67	0.15
Gender	−0.13*	0.07				
Avg_pedc	0.10	0.02				
$\sigma^2_{\theta^{(s)}}$	0.88	0.23				
$\sigma^2_{\theta^{(c)}}$	1.04	0.39				
#pars	77					
LL	−15,605.54					
VPC(s)	0.17					
VPC(c)	0.20					

Note. Gender = student gender (boys = 0, girls = 1); Avg_pedc = average parents' education centered. MLM = multilevel modeling; 2PL = two-parameter logistic; LL = log likelihood; VPC = variance partitioning coefficient.

*p < .05.

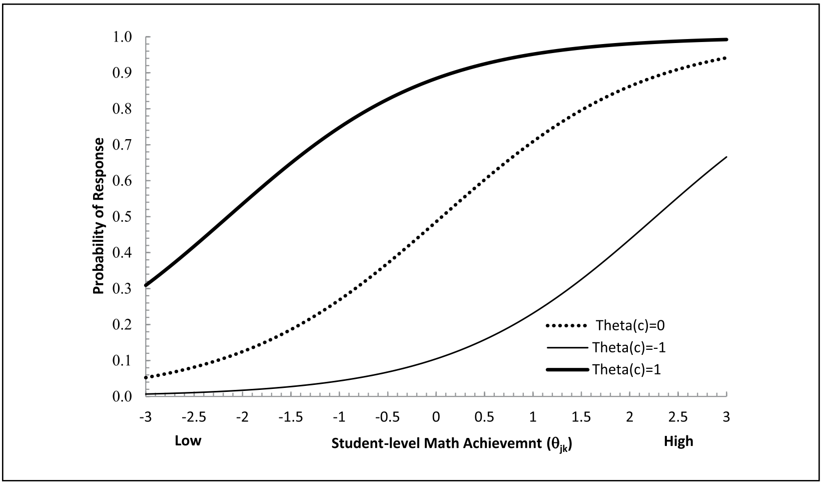

Figure 2. Student-level item response function conditioned at given class levels for Item 18 on the 25-item math achievement test fit by the MLM-2PL model.
Note. Horizontal axis represents the level of the latent trait (which has a standard normal distribution by construction), and vertical axis that measures the probability of a correct response at a specified latent trait level. Theta(c) = class-level math achievement fixed at a given level. To create the CIRFs, the model was identified by setting the variances of the latent traits to 1. MLM = multilevel modeling; 2PL = two-parameter logistic; CIRFs = conditional item response functions.

conditional item response functions (CIRFs). To create the CIRFs for any item, it is convenient to use Equation 6 and the MLIRT estimates where the model is identified by setting the variances of the latent trait to 1. The estimates of the model with variance of the random term fix equal to 1 can be approximated by multiplying each loading estimate in the MLM-2PL model in Table 5 by the standard deviation of the respective random term (Rabe-Hesketh, Skrondal, & Pickles, 2004a, 2004b). A CIRF that is built up conditioning on the student level (fixing θ_{jk} at a critical point and allowing θ_k free to vary) of the latent trait represents the probability to provide a correct response would vary across classes if they were composed of students with equal achievement in math for a given critical point. On the contrary, a CIRF that is built up conditioned on the class level (fixing θ_k at a critical point and allowing θ_{jk} free to vary) of the latent trait describes the probability of responding correctly to an item if students would attend classes which provide them the same contribution to their overall math achievement (i.e., classes with equal value of θ_k). The CIRFs for item 18 at student level with θ_k set equal to -1, 0, and 1 are plotted in Figure 2, whereas the CIRFs at class

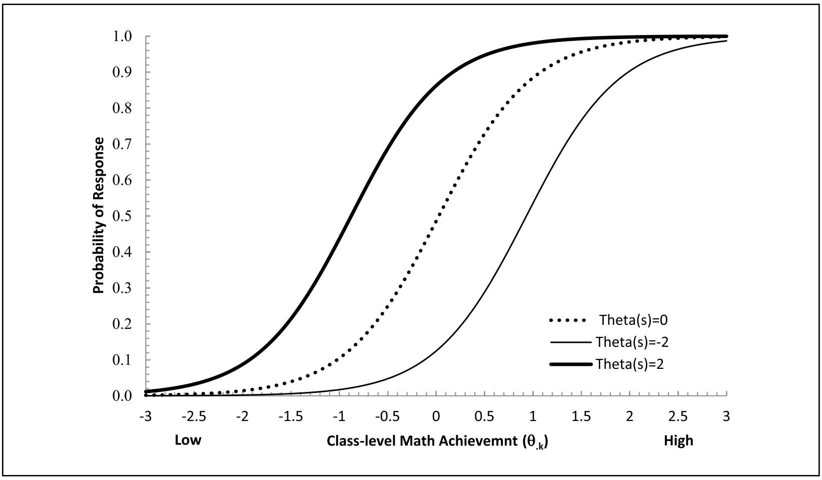

Figure 3. Class-level item response function conditioned at given student levels for Item 18 on the 25-item math achievement test fit by the MLM-2PL model.
Note. Horizontal axis represents the level of the latent trait (which has a standard normal distribution by construction), and vertical axis that measures the probability of a correct response at a specified latent trait level. Theta(s) = student-level math achievement fixed at a given level. To create the CIRFs the model was identified by setting the variances of the latent traits to 1. MLM = multilevel modeling; 2PL = two-parameter logistic; CIRFs = conditional item response functions.

level with θ_{jk} set equal to −2, 0, and 2 are plotted in Figure 3. An inspection of Figures 2 and 3 clearly shows how Item 18 has greater discrimination power at the class level ($\lambda_{18}^{(c)} = 2.09$; see Figure 3) versus the student level ($\lambda_{18}^{(s)} = 0.95$; see Figure 2) given the steeper CIRFs. In order to show the implications of allowing the loadings to vary at student level and class level and to highlight practical differences across MLIRT models with different loading constraints, we plot in Figure 4 the CIRFs at student level (fixing $\theta^{(c)} = 0$; plots a, c, and e in Figure 4) and class level (fixing $\theta^{(s)} = 0$; plots b, d, and f in Figure 4) of the MLM-Rasch, MLM-2PLC, and MLM-2PL model for items m2, m10, m11, m18. Comparing the MLM-2PL (plots e and f) versus the MLM-Rasch (plots a and b) shows that the latter model inflates the steepness of item m2 loading and deflates the steepness of item m18 loading at the class level. Thus, the MLM-Rasch model provides misleading information about the contribution of the items to the class-level latent trait values. A similar pattern is observed with MLM-2PLC model (plots c and d) versus the MLM-2PL model (plots e and f).

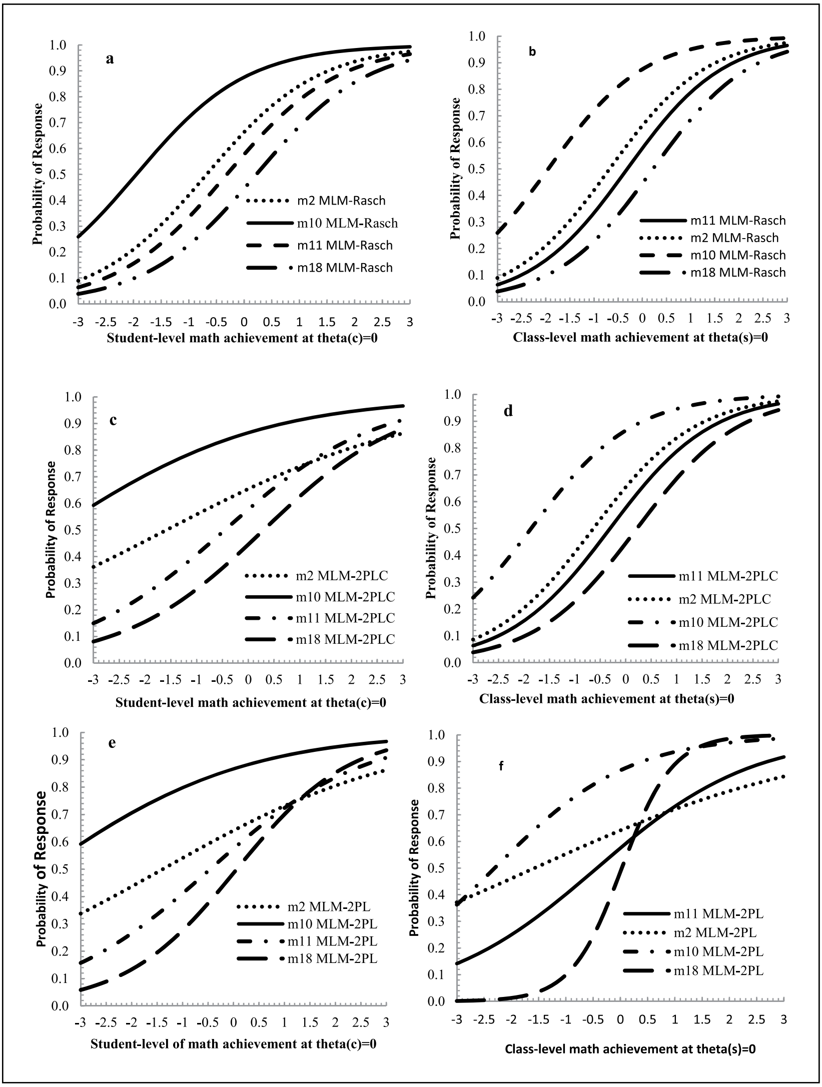

Figure 4. IRFs at student level (setting theta(c) = 0) and at class level (setting theta(s) = 0) for MLM-Rasch (a, b), MLM-2PLC (c, d), and MLM-2PL (e, f).

Note. The horizontal axis represents the level of the latent traits (which have a standard normal distribution by construction) at student level and class level, and vertical axis measures the probability of choosing or endorsing the category correctly at a specified latent trait level. IRFs = item response functions; MLM = multilevel modeling; 2PLC = two-parameter logistic constrained; MLM = multilevel modeling; 2PL = two-parameter logistic.

MLIRT information and reliability. Within the framework of classical test theory (CTT), reliability is the consistency of a group of persons' scores within one administration of a given multi-item instrument. A constant value is given to represent the degree of reliability associated with any raw score on a given instrument. Within this framework, a constant standard error of measurement value can be calculated to represent the precision or degree of error associated with any raw score on a given instrument.

Similarly, a single reliability index can be found within an IRT framework. A single IRT reliability index reflects, in general, the average or marginal reliability across the latent trait continuum (De Ayala, 2009, p. 177). Moreover, this IRT reliability index does not necessary reflect the amount of reliability in scores at the group average. For instance, an instrument may provide maximum reliability at $\theta = -1$ and less elsewhere, but the group average latent trait score could be above the location of maximal reliability. Thus, the IRT reliability index could over- or under-estimate the reliability reflected in the group average trait score. For this reason, it is not recommended to report a single IRT reliability index unless the total information function (TIF) is uniform (De Ayala, 2009). Information in IRT terms is similar to the concept of reliability, but information is allowed to vary based on the location along the latent trait continuum. As such, it is important and useful for researchers conducting MLIRT analyses to know the degree of conditional information at each level of analysis (e.g., class level, student level) and when appropriate the amount of information associated with various discrete and continuous covariates used within an MLIRT framework to explain or predict differences on the latent trait. We can use this information to identify the amount of reliable information at various points along the latent trait continuum for each discrete and continuous covariate.

TIF is a function of the cumulative information available in each item and as such each item has its own item information function (IIF). Drawing from classical IRT wherein item information at a given location is

$$\mathrm{IF}_i\left(\theta\right) = \frac{\left(P_i'\right)^2}{P_i\left(1-P_i\right)} = \lambda_i^2\, P_i\left(1-P_i\right), \tag{9}$$

where P_i (as defined in Equation 3) is the probability of a correct response for item *i* and $P_i^{'}$ is its first derivative with respect to the student parameter θ.

We can generate the IIF at the student level by deriving P_i (as defined in Equation 6) with respect to the person parameter and is expressed as

$$\mathrm{IIF}_i^{(s)}\left(\theta^{(s)}, \theta^{(c)}\right) = \frac{\left(P_i\left(\theta^{(s)}\right)'\right)^2}{P_i\left(1-P_i\right)} = \lambda_i^{(s)2}\, P_i\left(1-P_i\right). \tag{10}$$

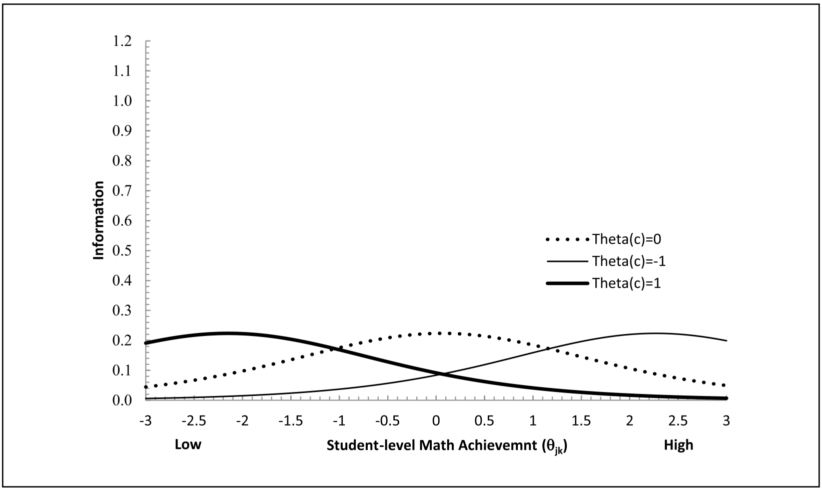

Figure 5. Student-level item information function conditioned at given class levels for Item 18 on the 25-item math achievement test fit by the MLM-2PL model.
Note. Horizontal axis represents the level of the latent trait (which has a standard normal distribution by construction), and the vertical axis measures the amount of student-level information at a specified latent trait level. Theta(c) = class-level math achievement fixed at a given level. To create the information functions the model was identified by setting the variances of the latent traits to 1. MLM = multilevel modeling; 2PL = two-parameter logistic.

Similarly, an IIF can be derived at the class level by deriving P_i with respect to the class parameter and is expressed as

$$\mathrm{IIF}_i^{(c)}\left(\theta^{(s)},\theta^{(c)}\right)=\frac{\left(P_i\left(\theta^{(c)}\right)'\right)^2}{P_i\left(1-P_i\right)}=\lambda_i^{(c)2}\,P_i\left(1-P_i\right). \tag{11}$$

The IIFs for Item 18 at student level with θ_k set equal to −1, 0, and 1 are plotted in Figure 5, whereas the IIFs at class level with θ_{jk} set equal to −2, 0, and 2 are plotted in Figure 6. An inspection of Figures 5 and 6 clearly shows how Item 18 has more information at the class level (see Figure 6) versus the student level (see Figure 5) given the higher peaks of the IIFs. An inspection of IIFs in Figure 5 shows that when conditioned at a class level of math achievement, maximal student-level information occurs at different points along the student-level latent continuum. For instance, given $\theta^{(c)} = 1$, the maximal student-level information occurs around $\theta^{(s)}$ −2.5, whereas when $\theta^{(c)} = -1$, maximal student-level information occurs around 2.5. The IIFs at student level and class level provide information about the contribution of each item for assessing students and class achievement. The inspection of the IIFs

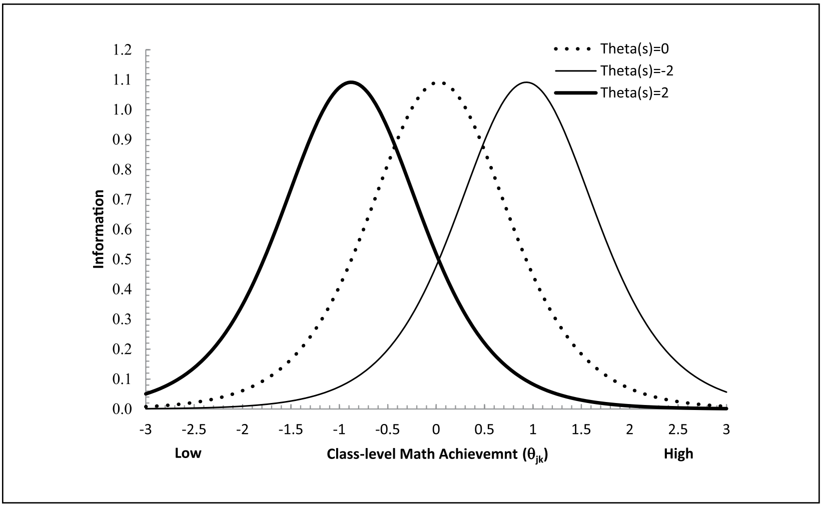

Figure 6. Class-level item information function conditioned at given student levels for Item 18 on the 25-item math achievement test fit by the MLM-2PL model.
Note. Horizontal axis represents the level of the latent trait (which has a standard normal distribution by construction), and the vertical axis measures the amount of class-level information at a specified latent trait level. Theta(c) = class-level math achievement fixed at a given level. To create the information functions the model was identified by setting the variances of the latent traits to 1. MLM = multilevel modeling; 2PL = two-parameter logistic.

at both levels helps to identify the amount of information each item provides to detect differences in mathematics achievement across students and classes. In practice, IIFs at both levels can be used to identify items that are helping to build up a test information function and which are not contributing much and need to be tossed or refined. Plus, the IIFs can be used to identify exemplary items that behave well at both levels for writing future items.

Broadly speaking, when class-level discrimination is zero, the IIFs conditioned at class level will be located at the same point, but when the ratio between class-level to student-level loading increases, the location of the IIFs will depart more and look similar to IIFs plotted in Figure 4. Interestingly, when student-level discrimination is zero, the IIFs conditioned at student level will be located at the same point, but when the ratio between student level to class level increases, the location of the IIFs will depart more.

The related TIFs are

$$\mathrm{TIF}_{..}^{(s)}\left(\theta^{(s)},\theta^{(c)}\right)=\sum\nolimits_{i}\mathrm{IIF}_{i}^{(s)}\left(\theta^{(s)}\theta^{(c)}\right) \tag{12}$$

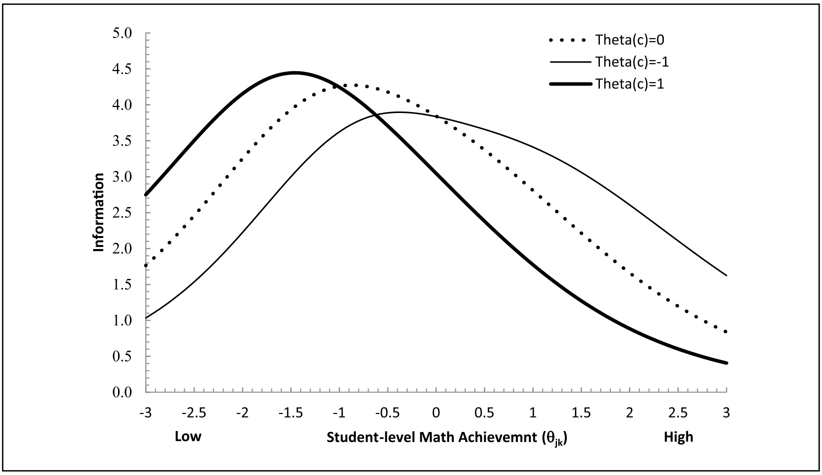

Figure 7. Student-level total information function conditioned at given class levels on the 25-item math achievement test fit by the MLM-2PL model.

Note. Horizontal axis represents the level of the latent trait (which has a standard normal distribution by construction), and the vertical axis measures the amount of student-level information at a specified latent trait level. Theta(c) = class-level math achievement fixed at a given level. To create the total information functions, the model was identified by setting the variances of the latent traits to 1. MLM = multilevel modeling; 2PL = two-parameter logistic.

and

$$\text{TIF}_{..}^{(c)}\left(\theta^{(s)},\theta^{(c)}\right)=\sum\nolimits_i \text{IIF}_i^{(c)}\left(\theta^{(s)}\theta^{(c)}\right). \tag{13}$$

Figure 7 shows the student-level TIFs conditioned at three points at the class level (−1, 0, and 1). Similar to the trends observed for the student-level IIFs, the shape of the TIFs show that when class-level achievement is high (1), the test is more informative for lower level students whereas when the class-level achievement is low (−1), the test is more informative for higher level students. Figure 8 shows the class-level TIFs conditioned at three points at the student level (−2, 0, and 2). Similar to the trends observed for the class-level IIFs, the shape of the TIFs show that when student-level achievement is high (2), the test tends to provide more information at lower points at the class level, whereas when student-level achievement is low (−2), the test tends to provide more information at higher points at the class level. The TIFs at student level and class level provide information about how well the test precisely assesses students and class achievement. The inspection of the TIFs at both levels helps to identify the amount of information a test provides to detect

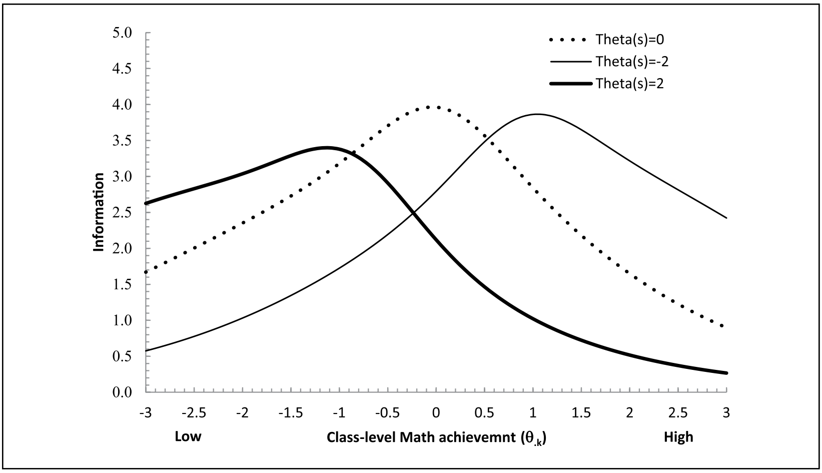

Figure 8. Class-level total information function conditioned at given student levels on the 25-item math achievement test fit by the MLM-2PL model.

Note. Horizontal axis represents the level of the latent trait (which has a standard normal distribution by construction), and the vertical axis measures the amount of class-level information at a specified latent trait level. Theta(s) = student-level math achievement fixed at a given level. To create the total information functions, the model was identified by setting the variances of the latent traits to 1. MLM = multilevel modeling; 2PL = two-parameter logistic.

differences in mathematics achievement across students and classes. In practice, TIFs at both levels can be used to identify how well a test is measuring points along the latent trait continuum and if more items are needed to better measure points on this continuum at any given level of the data structure.

Moreover, if it was important to report a single value to capture the precision of the entire range of the latent trait, where the TIF is uniform, a marginal reliability coefficient can be estimated by extending Green, Bock, Humphreys, Linn, and Rechase (1984) concept of marginal reliability to MLIRT. Marginal reliability at the student level $\cong 1 - 1/\text{Information}(\theta^{[s]}|\theta^{[c]})$ and at the class level $\cong 1 - 1/\text{Information}(\theta^{[c]}|\theta^{[s]})$. For instance, if we fix $\theta^{(c)} = 0$, then marginal reliability at the student level is $1 - 1 / 4.3 = .77$ at $\theta^{(s)} = -0.8$. Similarly, if we fix $\theta^{(s)} = 2$, then marginal reliability at the class level is $1 - 1 / 3.4 = .71$ at $\theta^{(c)} = -1.1$.

Explanatory MLIRT. Once the MLIRT model has been built-up for descriptive purposes, researchers can attempt to explain the variability in the value of the latent trait (math achievement) at both student level and class level by including relevant covariates. The latent traits can be related to student and class characteristics expressed as

$$P(m_{ijk}=1\,|\,\theta,\alpha,\lambda,\beta_s,\beta_c)=\frac{e^{\alpha_i x_i+\lambda_i^{(s)}\left(\sum_{s=1}^{S}\beta_s z_{jks}+\tilde{\theta}_{jk}\right)+\lambda_i^{(\mathrm{c})}\left(\sum_{c=1}^{C}\beta_c w_{.kc}+\tilde{\theta}_k\right)}}{1+e^{\alpha_i x_i+\lambda_i^{(s)}\left(\sum_{s=1}^{S}\beta_s z_{jks}+\tilde{\theta}_{jk}\right)+\lambda_i^{(c)}\left(\sum_{c=1}^{C}\beta_c w_{.kc}+\tilde{\theta}_k\right)}}, \quad (14)$$

where $\tilde{\theta}_{jk}$ is the adjusted student-level value of the latent trait after having removed the effect of student-level covariates *z* (S = number of student-level covariates), whereas $\tilde{\theta}_k$ is the adjusted class-level value of the latent trait after having removed the effect of the class-level covariates *w* (C = number of class-level covariates). The model expressed in Equation 14 is an extension of a *latent regression model* (Zwinderman, 1991). The model in Equation 14 includes two random terms. The variability of the random term $\tilde{\theta}_{jk}$ ($N\left[0,\sigma^2_{\left(\tilde{\theta}^{(s)}\right)}\right]$) captures the between-students within-classes (student level) variance after having accounted for any student-level covariates, whereas the random term $\tilde{\theta}_k$ ($N\left[0,\sigma^2_{\left(\tilde{\theta}^{(c)}\right)}\right]$) captures the between-class variance after adjusting for any class-level covariates.

Explanatory MLIRT results. Using Equation 14, we can add covariates of interest and test the following research question: What is the value added by fitting an explanatory MLIRT model with covariates at the student level to the analysis of students' achievement in math? Specifically, in statistical terms, what is the nature and significance of the slope between student-level covariates (gender and parents education) as they predict scores on the latent variable (math)?

We answered this question by specifying an MLM-2PL model with covariates (MLM-2PL WC) which models the variability in the latent trait at student level as a function of gender (gender) and average level (years) of education of the students' parents centered (avg_pedc) at the grand mean (M = 10.74, SD = 3.1). Observed values for avg_ped ranged from 5 to 17. A comparison of the MLM-2PL WC model to the model without covariates (MLM-2PL) shows it had better fit to the sample data. Specifically, the LRT was −2(−15,648.45 – −15,605.54) = 85.82 with df = 77 – 75 = 2, which has an associated p value < .001. Therefore, the LRT shows that including the student-level covariates improves the explanation of the variability in item responses.

Table 5 summarizes the results for the MLM-2PL WC model. An inspection of the student-level covariates shows gender and avg_pedc have significant slopes with the latent trait at the student level. Specifically, female students have on average lower achievement in math than male students, whereas more educated parents have on average a positive slope with the latent trait. The coefficient parameters highlight that the 577 female students have an average achievement in math that is 0.13 lower than the 564 males conditional upon the parents' level of education. The 30 male students with avg_ped = 17 have an overall achievement that is 0.40 higher than the 101 male students who have avg_ped = 13 is about 0.63 higher than male students whose parents have a level of education equal to the average (avg_ped: *M* = 10.73). If we compare two students with the highest and lowest profile of both explanatory predictors in terms of change to score correctly (i.e., specifically the 30 "Male students with avg_ped = 17" and the 11 "Female students with avg_ped = 5"), the former have an achievement in math test that is on average about 1.33 higher than the latter. Translating this difference in terms of standard deviations from the average math achievement on the latent variable (that is equal to zero), this means that the two students' profiles record on average an expected difference in the achievement equal to 1.55 *z* scores. This is akin to an effect size of 1.55 standard deviation differences between these groups. This can happen because the latent trait metric is standardized by default to mean of 0 and unit variance. It is worth highlighting that the adjustment for the two students' covariates shift up the lower bound of the range of variation of the latent trait (math) and this clearly arises by comparing the range of variation of prediction of the latent trait at the student level [−2.60 to 2.49] with the range of variation of the adjusted student level [−2.39 to 2.48]. These results suggest that the socio-cultural characteristics of the family, summarized by the variable avg_ped, partially explain differences in students' achievement in math achievement. This is highlighted also by comparing the variability of the random terms of models MLM-2PL versus MLM-2PL WC, from which it can be stated that the adjustment for covariates reduces the unexplained variability of the random terms at the student level and there is an increase of the variability at the class level. Thus, the omission of relevant students' covariates leads to an overestimation of individual students' differences. This means that the adjustment for covariates allows us to better understand how much of the divergences in math achievement are due to the two student-level characteristics, as it clearly arises by a comparison between the values of the variances of the two random terms, namely $\sigma^2_{\tilde{\theta}^w} = 0.88$ and $\sigma^2_{\tilde{\theta}^B} = 1.03$. The two student-level covariates considered in the analysis account for approximately 3% of the variability in math achievement (see Table 4, MLM-2PL VPC(s) = 0.20 and Table 5, MLM-2PL WC VPC(s) = 0.17). The MLM-2PLWC model highlights how

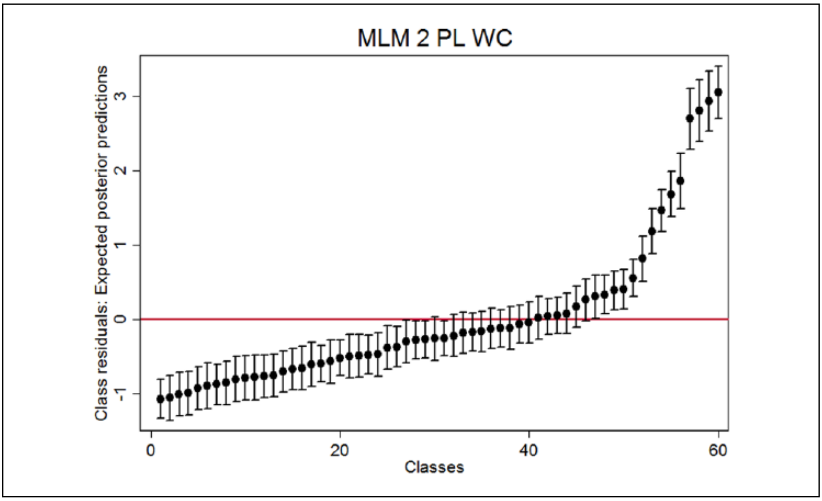

Figure 9. Expected posterior predictions at class level with 95% CI.
Note. MLM-2PL WC = multilevel modeling two-parameter logistic with covariates; CI = confidence interval.

accounting for students' covariates increases the shared amount of unexplained variability due to differences across classes.

Table 5 provides the item parameters from the MLM-2PLWC model. If we rank the items according to the value of the item intercept, we observe that items "$m_{24,}$ m_1, and m_{10}" are the ones for which the probability to answer correctly is highest, whereas items "m_9, m_{17}, and m_7" are the ones for which the probability to answer correctly is lowest. It is worth highlighting that the adjustment for student-level covariates did not substantively change the variation of the item intercept parameters in the MLM-2PL model (see Table 4) versus the MLM-2PL WC model (see Table 5).

The within-class between-student loadings (Level 2) and the between-classes factor loadings (Level 3) ranged from 0 to 1.75 and 0.37 to 2.03, respectively (see Table 5). The correlation between the loadings at class level and the loadings at student level was low ($\hat{\rho}$ = 0.06), which indicates that the same item may have high discrimination power at student level and lower at class level or vice-versa. In essence, item loadings are not always the same at all levels of analysis. For instance, item $m_{23,}$ which is the item with the greatest power to discriminate between students $(\lambda_{23}^{(s)} = 1.75)$, but has a lower power to discriminate between classes $(\lambda_{23}^{(c)} = 0.72)$.

Predictions of the latent traits. Once item parameters and variances of the random terms (if requested) are estimated, with or without covariates, random person (student) and cluster (class) parameters are estimated (predicted) in a second step (see Muthén & Muthén, 1998-2015; Skrondal & Rabe-Hesketh, 2004, 2009; StataCorp, 2013). In general, the information or details about any given latent student and class parameters is found by taking the resulting predictions for $\tilde{\theta}_{jk}$ and $\tilde{\theta}_k$ with the respective measures of uncertainty (which can be interpreted as a standard error). For instance, student-level $\tilde{\theta}_{jk}$ is an adjusted indicator of student math achievement estimated by subtracting the effect of factors z on the parameter θ_{jk}. The literature on factor models refers to these measures as "factor scores," whereas the literature on multilevel models calls them "empirical Bayes predictions," "posterior prediction," or "higher level residuals" (Rabe-Hesketh et al., 2004b; Skrondal & Rabe-Hesketh, 2009). Finally, comparisons of each class to the average class-level achievement can be done using the predicted random term at the class-level θ_k with the respective measure of uncertainty SD_k^{θ}. Figure 9 plots classes according to their latent level of math achievement based on 95% CIs. If we want to make a paired-comparison between two classes using a 95% CI, then we need to construct the caterpillar plot by taking the standard 1.96 /√2 (Goldstein & Healy, 1995; Goldstein & Spiegelhalter, 1996).

Estimation methods. The likelihood function of IRT models which assumes a continuous distribution of the random effects (as the likelihood of the models just described) cannot be maximized using analytical methods because it does not have a closed form solution. The integrals involved in the functions need to be first approximated by sums before to start the maximization process. The Stata procedures GLLAMM and SEM and the program M*plus* have the option of using a maximum likelihood estimation method that marginalizes the likelihood function by integrating out the random terms using a numerical quadrature method. The numerical integrated marginal likelihood is then maximized using the Newthon-Raphson algorithm. In all models fitted in this article with Stata, we adopted marginal maximum likelihood (MML) with adaptive quadrature, which ensures good estimates of the parameters compared with other quadrature methods (e.g., Gauss Hermite is less accurate also in cases where standard quadrature work poorly, see for more details Rabe-Hesketh, Skrondal, & Pickles, 2004a). Furthermore, adaptive quadrature works faster and needs fewer quadrature points [nodes]. In flexMIRT, we used both Bock-Aitkin EM (BAEM) algorithm (Bock & Aitkin, 1981) and the Metropolis Hastings-Robbins Monro (MH-RM; see Cai, 2010) algorithm for estimating all descriptive IRT and MLIRT models. In flexMIRT, MLIRT models can be estimated using also the default BAEM;

however, it can take noticeable time to converge, whereas the MLIRT WC models can be estimated only with the MH-EM algorithm. In our experiences, the BAEM in flexMIRT takes considerably longer with MLIRT models estimating more than one latent trait dimension, but are efficient with MLIRT without covariates as considered herein.

The estimates of the MH-EM algorithm in flexMIRT for Rasch and 2PL models almost match (Cai, 2009, 2010) with the results which rely on the other marginal maximum likelihood algorithms available in flexMIRT, Stata and M*plus*, whereas the estimates of the MLIRT models with the MH-EM show some slight differences with respect to the results observed with the other estimation methods. In our experience the biggest divergences between the MLM IRT model estimates obtained with marginal maximum likelihood algorithms in Stata, M*plus* and flexMIRT and the MH-EM results are observed in the estimation of the variance of the class-level random term. Specifically, with respect to the algorithms which rely on the likelihood, the MH-EM algorithm increases the estimates of the class-level random term in the MLM-Rasch and downward the class-level random term in the MLM-2PL and MLM-2PL WC.

In general, Stata, M*plus*, and flexMIRT provide similar results for the variety of models considered in this article, but the main drawbacks of Stata GLLAMM are related to the computational time required to approximate the likelihood which increases as a function of the number of latent variables, the number of levels, the number of quadrature points, the number of the observations in the data set, and the square of the number of parameters involved in the estimation. It is interesting to highlight that for models that can be specified with both Stata routines, it is more convenient to use the Stata program SEM that is more efficient in terms of computational time. Specifically, Stata program SEM is faster than GLAMM, but for more complex models (e.g., multidimensional MLIRT with more than three levels) GLAMM can take considerably longer to converge (Rabe-Hesketh et al., 2004b; StataCorp, 2013). From our experiences, this efficiency is also true when using M*plus* and flexMIRT. For the models considered herein, the most efficient (fastest) software, in rank order, were flexMIRT, M*plus*, Stata program SEM, and Stata program GLLAMM.

We presented the results using the Stata program GLAMM because it is the only routine that estimates MLIRT using a data set in long format within a generalized multilevel or three-level multilevel model, whereas the Stata program SEM and M*plus* estimates MLIRT models using a data set in wide format within a two-level multilevel multivariate model. Appendix B provides a summary comparison of some of the software capabilities and features with respect to the MLIRT models for binary responses considered in this article. In general, Stata programs GLAMM and SEM, M*plus*, and flexMIRT have similar capabilities and limitations, except Stata programs GLAMM and SEM allow users to model more than three levels of data.

Discussion

In this article, we have provided the advantages afforded to researchers when MLIRT is used for descriptive and explanatory purposes. We provide details for applied researchers wanting to conduct MLIRT analyses for developing an instrument using IRT and for testing theories and research questions within the MLIRT framework. This article shows how to test if an MLIRT analysis is necessary and showcases how to create item response, information, and total information functions within a multilevel context. This article also shows the potential of explanatory MLIRT modeling and explains how to interpret the results. We also highlight the meaning of the student-level and class-level latent traits and how to interpret these measures when covariate are included. Furthermore, this article shows how to interpret the Level 2 and Level 3 loadings. The use of this model can be particularly fruitful in the framework of early adolescence because the data are usually collected in a nested data structure. We recommend that MLIRT be used whenever the focus of a study is on (a) selecting items to build up an informative instrument at class- and/or student level(s); (b) measuring the latent traits at individual- and class level(s); (c) adjusting individual or class scores for relevant covariates at student- and/or class level(s); (d) making comparative evaluations across students and classes. Naturally, MLIRT can be extended to contexts that do not consider students within classes (e.g., teachers in schools, athletes in teams). Researchers also need to be aware that the validity of the conclusions made with MLIRT depends on having adequate sample sizes at each level of interest based on the research questions under consideration. Although we do not check assumptions (i.e., dimensionality, LI, monotonicity) specific to IRT within the MLM context, we recommend that this be checked as well when doing MLIRT. We recommend that future directions on MLIRT should work on how to do MLIRT with polytomous models, multidimensional models, higher level covariates, cross-level interactions, mediation, latent predictors, handling missing data, longitudinal data structures, develop and evaluate methods for testing MLIRT assumptions, and equations and software programs for conducting power and sample-size determination. Such future work would aid researchers wanting to do MLIRT and how to determine adequate sample sizes. If future research focuses on the above mentioned recommendations, then more researchers will be able to have access to information about MLIRT, on how to do it in a variety of contexts, and how to determine samples needed for doing MLIRT analyses. Finally, in the References section, we identify for readers wanting more technical or methodological readings on this topic to seek out those with an *, whereas those wanting some quality applied examples to read those **. We hope that this article helps applied researchers to appreciate and understand MLIRT and can move the field of early adolescence and MLIRT forward.

Appendix A

Table A1. Wide Form "IMVALSIMATH.CSV."

CLASS	ID	M1	M2	M3	M4	M5
1	1	1	1	1	1	0
1	2	1	1	0	1	1
1	3	1	1	0	1	1
1	4	1	1	0	0	0

Table A2. Long Form "INVALSIMATHLONG.CSV."

CLASS	ID	m	item
1	1	1	1
1	1	1	2
1	1	1	3
1	1	1	4
1	1	0	5
1	2	1	1
1	2	1	2
1	2	0	3
1	2	1	4
1	2	1	5
1	3	1	1
1	3	1	2
1	3	0	3
1	3	1	4
1	3	1	5
1	4	1	1
1	4	1	2
1	4	0	3
1	4	0	4
1	4	0	5

Table A3. Long Form "INVALSIMATHLONG.CSV" Plus Dummy Indicators.

CLASS	ID	m	item	x1	x2	x3	x4	x5
1	1	1	1	1	0	0	0	0
1	1	1	2	0	1	0	0	0
1	1	1	3	0	0	1	0	0
1	1	1	4	0	0	0	1	0
1	1	0	5	0	0	0	0	1
1	2	1	1	1	0	0	0	0
1	2	1	2	0	1	0	0	0
1	2	0	3	0	0	1	0	0
1	2	1	4	0	0	0	1	0
1	2	1	5	0	0	0	0	1
1	3	1	1	1	0	0	0	0
1	3	1	2	0	1	0	0	0
1	3	0	3	0	0	1	0	0
1	3	1	4	0	0	0	1	0
1	3	1	5	0	0	0	0	1
1	4	1	1	1	0	0	0	0
1	4	1	2	0	1	0	0	0
1	4	0	3	0	0	1	0	0
1	4	0	4	0	0	0	1	0
1	4	0	5	0	0	0	0	1

Appendix B

Comparison of Software Capabilities and Features With Respect to MLIRT Models for Binary Item Responses Considered in This Article.

Capabilities	Stata 13 (GLLAMM)	Stata 13 (SEM)	M*plus*7.4	flexMIRT3.0
MLM-Rasch	Y	Y	Y	Y
MLM-2PLC	Y	Y	Y	Y
MLM-2PL	Y	Y	Y	Y
LI Test	N	N	N	N
Item-Level Model-Fit test	N	N	N	N
IIF for MLIRT	N	N	N	N
TIF for MLIRT	N	N	N	N
Descriptive	Y	Y	Y	Y
Explanatory	Y	Y	Y	Y

(continued)

Appendix B (continued)

Capabilities	Stata 13 (GLLAMM)	Stata 13 (SEM)	M*plus*7.4	flexMIRT3.0
Data Format (long, wide)	Long	Wide	Wide	Wide
Maximum number of levels that can be modeled	More than 3	More than 3	3	3
Binary predictor	Y	Y	Y	Y
Continuous Predictor	Y	Y	Y	Y
Bock-Aitkin EM (BAEM) algorithm and marginal maximum likelihood estimation method	Y	Y	Y	Y
Metropolis Hastings-Robbins Monro (MH-RM) algorithm	N	N	N	Y
Treatment of levels in software				
Level 1	Item responses	Student	Student	Class
Level 2	Student	Class	Class	Student
Level 3	Class	—	—	—
Parametrization of the 2PL model: $e^z/(1 + e^z)$	$z=\alpha_i+\lambda_i\theta_j$	$z=\alpha_i+\lambda_i\theta_j$	$z=-\alpha_i+\lambda_i\theta_j$	$z=\alpha_i+\lambda_i\theta_j;\alpha=c$

Note. Binary or continuous predictors can only be included in flexMIRT using the Metropolis Hastings-Robbins Monro (MH-RM) algorithm, which is an approximation method. The loadings of the model with variance of the random terms fix equal to 1 can be approximated by multiplying the loadings of the models with unconstrained variance of the random terms by the standard deviation of the respective random terms. GLLAMM = generalized linear latent and mixed models; flexMIRT = flexible multilevel multidimensional item analysis and test scoring; MLIRT = multilevel item response theory; SEM = structural equation modeling; MLM = multilevel modeling; MLM-2PLC = multilevel modeling two-parameter logistic constrained; MLM-2PL = multilevel modeling two-parameter logistic; IIF = item information function; TIF = total information function; EM = expectation maximization; BAEM = Bock-Aitkin EM; MH-RM = Metropolis Hastings-Robbins Monro.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions, and Caihong Li and Abbey Love for their careful review of early versions of this manuscript.

Authors' Note

Order of authorship is alphabetical.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. In the Stata program, SEM adopts the parametrization specified in Equation 2, whereas in M*plus*, parametrization of the intercept parameter α_i enters with a negative sign in Equation 2, thus the relation between easiness parameter and location parameter is $b_i = \alpha_i \big/ \lambda_1$.
2. In a binary model, the variance of the Level 1 residual variance is fixed and thus it does not decrease when relevant covariates are introduced into the analysis. If a multilevel logistic regression model is specified, an additional source of variability is introduced at Level 2 and the Level 1 variance cannot decrease. This stretches the scale of the response and as a consequence all the fixed-effect coefficients will be greater in magnitude. For this reason, the introduction of covariates tends to increases the variance explained at higher levels (for details, see Steele, 2009).

References

References marked with an asterisk indicate studies with a more technical or methodological focus on this topic.

References marked with a double asterick indicate exemplar applied studies on this topic.

*Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Browne, W. J., Golalizadeh Lahi, M., & Parker, R. M. A. (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim software package*. Bristol, UK: University of Bristol.

Cai, L. (2009). High-dimensional exploratory item factor analysis by A Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33-57.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307-335.

Cai, L. (2015). flexMIRT®: Flexible Multilevel Multidimensional Item Analysis and Test Scoring (Version 3.03) [Computer software]. Chapel Hill, NC. Vector Psychometric Group.

Cho, S.-J., & Suh, Y. (2012). Bayesian analysis of item response models using WinBUGS 1.4.3. *Applied Psychological Measurement*, *36*, 147-148.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Ayala, R. J. (2013). Factor analysis with categorical indicators: Item response theory. In Y. Petcscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 208-242). New York, NY: Routledge.

*De Boeck, P., & Wilson M. (Eds.). (2004). *Item response models: A generalized linear and nonlinear approach* (Statistics for social and behavioral sciences). New York, NY: Springer.

**Draper, D., & Gittoes, M. (2004). Statistical analysis of performance indicators in UK higher education. *Journal of the Royal Statistical Society, Series A*, *167*, 449-474.

Edelen, M. O., & Reeve, B. B. (2007). Applying item theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5-18. doi:10.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fisher, G. H., & Molenaar, I. W. (1995). *Rasch models, foundations, recent developments, and applications*. New York, NY: Springer-Verlag.

Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research*, *103*, 231-265.

*Fox, J.-P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, *20*(5), 1-16.

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

Glaser, D., & Hastings, R. H. (2011). An introduction to multilevel modeling for anesthesiologists. *Statistical Ground Rounds*, *113*, 877-887.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A*, *158*, 175-177.

**Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of Royal Statistical Society A*, *159*, 385-443.

**Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society A*, *159*, 97-114.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Rechase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *32*, 347-360.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*, 385-410. doi:10.1111/j.1468-2958.2006.00281.x

Hox, J. J. (2010). *Multilevel analysis methods: Techniques and applications*. New York, NY: Routledge.

Jak, S., Oort, F., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 31-39.

Jeon, M., & Rabe-Hesketh, S. (2012). Profile likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behavioral Statistics, 37*, *4*, 518-542.

Jeon, M., & Rijmen, F. (2015). A modular approach for item response theory modeling with the R package flirt. *Behavioral Research Methods*. doi:10.3758/s13428-015-0606-z.

**Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*, 82-100.

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

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: SAGE.

Linn, R. L. (2009). The concept of validity in the context of NCLB. In R.W. Lissitz. (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 195-212). Charlotte, NC: IAP-Information Age Publishing.

Lord, F. M. (1952). *A theory of test scores* [Psychometric Monographs, No. 7]. http://psycnet.apa.org/psycinfo/1954-01886-001.

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

McDonald, R. P. (1999). *Test theory: A unified treatment*. New York, NY: Routledge.

McEldoon, K., Cho, S.-J., & Rittle-Johnson, B. (2012). *Measuring intervention effectiveness: The benefits of item response theory approach*. SREE fall 2012 conference abstract template. Retrieved from https://www.sree.org/conferences/2012f/program/downloads/abstracts/693.pdf

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267-316.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

**Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied measurement in education*, *16*, 223-243.

Peugh, J. L. (2014). Conducting three-level cross-sectional analyses. *The Journal of Early Adolescence*, *34*, 7-37.

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

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004b). *GLAMM manual* (U.C. Berkeley Division of Biostatistics Working Paper Series, 160). Berkeley: University of California.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydicke.

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

Ravand, H. (2015). Item response theory using hierarchical generalized linear models. *Practical Assessment Research & Evaluation*, *20*, 1-17.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of personality assessment*, *84*, 126-136.

Rijmen, F., & Briggs, D. (2004). Multiple person dimensions and latent item predictors. In P. De Boeck & M. Wilson (Eds), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 247-265). New York, NY: Springer.

*Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185-205.

Singer, J. D. (1998). Using SAS Proc Mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*, 323-355. doi:10.3102/10769986023004323

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variables modeling*. Boca Raton, FL: Chapman & Hall.

*Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society Series A*, *172*, 659-687.

Snijders, T. A. B. (2005). Power and sample Size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1570-1573). Chichester, UK: Wiley.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: SAGE.

StataCorp. (2013). *Stata statistical software: Release 13*. College Station, TX: StataCorp LP.

Steele, F. (2009). Multilevel models for binary responses- Concepts. *LEMMA VLE Module*, *7*, 1-41. Retrieved from http://www.bristol.ac.uk/cmm/learning/course.html

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*, 120-151. doi:10.1177/0272431613511332

Vector Psychometric Group. (2015). *FlexMIRT plotting manual*. Retrieved from http://www.vpgcentral.com/wp-content/uploads/2014/03/flexMIRTplotting-Manual.pdf

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145.

Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. Bovaird, K. Geisinger, & C. Buckendahl (Eds.) *High stakes testing in education—Science and practice in K-12 settings* (Festschrift to Barbara Plake) (pp. 177-190). Washington, DC: American Psychological Press.

Zwinderman, A. H. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, *56*, 589-600.

Author Biographies

Isabella Sulis is an assistant professor in the Department of Social Sciences and Institution at the University of Cagliari. Her research interests include multilevel analysis, item response theory, measurement model to build-up adjusted indicators, and missing data.a

Michael D. Toland is an associate professor in the Educational Psychology program in the Department of Educational, School, and Counseling Psychology at the University of Kentucky. His research interests include psychometrics, item response theory, factor analysis, scale development, multilevel modeling, and the realization of modern measurement and statistical methods in educational research.