Multilevel mixture item response theory models: an application in education testing

Jeroen K. Vermunt
Tilburg University, Department of Methodology and Statistics
Warandelaan 2
Tilburg 5037 AB, The Netherlands
E-mail: j.k.vermunt@uvt.nl

Introduction

Skrondal and Rabe-Hesketh (2004) proposed a generalized latent variable modeling framework integrating 1) factor analytic and random coefficient models, 2) models with discrete and continuous unobserved variables, and 3) hierarchical models with unobserved variables at different levels. This framework is implemented in their GLLAMM software package. In this paper I describe a strongly related framework that is implemented in the syntax version of the Latent GOLD software (Vermunt and Magidson, 2007). The most important extension compared to the GLLAMM approach is that it allows defining models with any combination of discrete and continuous latent variables at each level of the hierarchy. The approach is illustrated with a multilevel application in educational testing; that is, using a set of mathematics test items taken from pupils nested within schools. An item response theory model is constructed for the responses on the test items and the between-school differences in pupils' abilities and item difficulties is modeled using a discrete mixture distribution at the school level.

The generalized latent variable model

The generalized latent variable model contains four elements: 1) multivariate responses or *dependent variables* ($\mathbf{y}$), which may be binary, nominal, ordinal, continuous, counts, or any combination of these; 2) *latent variables* ($\boldsymbol{\nu}$), which may be discrete (nominal or ordinal), continuous, or combinations of these; 3) predictors or *independent variables* ($\mathbf{Z}$ and $\mathbf{W}$); and 4) nested or *multilevel observations* at L levels. Using the index k to denote an independent observation corresponding to the highest level of the hierarchy, the model can be formulated with the following two equations:

$$g[E(\mathbf{y}_k)] = \mathbf{Z}_k^{(1)}\boldsymbol{\beta} + \mathbf{W}_k^{(1)}\boldsymbol{\Lambda}^{(1)}\boldsymbol{\nu}_k \quad (1)$$

$$h[E(\boldsymbol{\nu}_k^{(\ell)})] = \mathbf{Z}_k^{(\ell)}\boldsymbol{\gamma}^{\ell} + \mathbf{W}_k^{(\ell)}\boldsymbol{\Lambda}^{(\ell)}\boldsymbol{\nu}_k^{(\ell^+)} \quad \text{for } \ell = 2, ..., L. \quad (2)$$

Here, $g[\cdot]$ and $h[\cdot]$ are link functions (identity, logit, log, etc.) which may differ across dependent variables and across latent variables and which typically depend on the scale type of the left hand variable. The free model parameters are the regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\gamma}$, as well as the residual (co)variances (or associations) between latent variables and between dependent variables. Note that $\boldsymbol{\nu}_k$ denotes the vector of latent variables of observation k at all levels, whereas $\boldsymbol{\nu}_k^{(\ell)}$ and $\boldsymbol{\nu}_k^{(\ell^+)}$ refer to the latent variables at level ℓ and ℓ and higher, respectively. Details on maximum likelihood estimation of these types of models using the EM algorithm can be found in Vermunt (2004).

In one aspect, the framework implemented in Latent GOLD is slightly less general than suggested by the two model equations: the structural equation model for the latent variables at level ℓ is only partially implemented. But in other aspects it is even more general than expressed in the above two equations, including that it allows specification of a Markovian structure for discrete latent variables at the lowest level, of interaction effects between latent variables, and of many different models for the residual (co)variances and associations.

It is important to note that the product term $\mathbf{W}_k^{(1)}\mathbf{\Lambda}^{(1)}$ in equation (1) is what yields the generalization of both the factor analytic and the random coefficient model. $\mathbf{\Lambda}^{(1)}$ is the factor loadings matrix of a factor analysis and $\mathbf{W}_k^{(1)}$ is the design matrix of a random coefficient model. This implies that by setting $\mathbf{W}_k^{(1)} = \mathbf{1} \otimes \mathbf{I}$ we obtain a factor analytic model and by setting $\mathbf{\Lambda}^{(1)} = \mathbf{I}$ we obtain a random coefficient model. The product $\mathbf{W}_k^{(1)}\mathbf{\Lambda}^{(1)}$ – which Skrondal and Rabe-Hesketh (2004) refer to as the structure matrix $\mathbf{\Lambda}_k^{(1)}$ – defines the generalized latent variable framework in which the effects of latent variables on responses may contain parameters, fixed terms, or products of these.

It should be noted that the latent variables $\boldsymbol{\nu}_k$ can be common factors in a factor analysis, random coefficients in a multilevel model, classes in a latent class model, or mixture components in a finite mixture model. In other words, the latent variables may be either discrete or continuous and may be used either to reveal structure (meaningful factors or clusters) or to correct for unobserved heterogeneity.

Overview of the special cases

Table 1. Nine-fold classification of possible models with latent variables at two levels

	Higher-level ν's		
Lower-level ν's	Continuous	Discrete	Combination
Continuous	A1	A2	A3
Discrete	B1	B2	B3
Combination	C1	C2	C3

Assuming two levels of latent variables and taking into account that the latent variables at each level may be continuous, discrete, or a combination of these, we obtain the nine-fold classification provided in Table 1. One of the special cases, in which both the lower- and higher-level latent variables are discrete (B2), is the hierarchical variant of the latent class model proposed by Vermunt (2003). Here, lower-level units (cases) are clustered based on their observed responses as in a standard latent class model, whereas higher-level units (groups) are clustered based on the likelihood of their members to be in one of the case-level clusters. Vermunt (2003) also proposed a multilevel latent class model with continuous random effects at the group level which belongs to category B1.

A1 contains both three-level regression models with continuous random effects and two-level factor analytic and item response theory (IRT) models, such as the multilevel IRT model proposed by Fox and Glas (2001). In a recent paper, Palardy and Vermunt (2007) used specification A3 for defining a multilevel extension of the mixture growth model. In the application described below, we use a type A2 model.

What is clear from the above table is that the presented framework yields a large number of options. With latent variables at three instead of two levels, the number of possible specifications increases from 9 to 27. It depends on the specific application which of the specifications should be selected; that is, whether it is more meaningful and/or practical to define the latent variables at a particular level to be continuous, discrete, or a combination of the two.

An illustrative application: a multilevel mixture IRT model

The application uses a data set collected by Doolaard (1999), and which was also used by Fox and Glas (2001) to illustrate their multilevel IRT model. More specifically, information is available on a 18-item math test taken from 2156 pupils belonging to 97 schools in the Netherlands. The aim of the analysis is twofold: measuring pupils' math abilities and assessing differences between school. The first aim involves building a single factor or IRT model for the 18 math items, while the second

aim involves introducing school-level random coefficients in the IRT model.

As far as the IRT model is concerned, two different models are considered: the two-parameter logistic (2-PL) model and the Rasch model, which is also referred to as the one-parameter logistic (1-PL) model. As in Fox and Glas's multilevel IRT model, we are interested in school differences in ability. Unlike Fox and Glas, we also want to know whether the items' functioning is the same across schools; that is, we want to perform what is usually referred to as an item bias analysis. This is feasible using a discrete finite mixture specification for the relevant school differences. The proposed multilevel mixture IRT model can, therefore, be seen as a practical method for detecting item bias in situations in which the number of groups is too large for a standard item bias analysis.

Let y_{ijk} denote the binary response on item i of pupil j in school k. Note that i, j, and k refer to a level-1, level-2, and level-3 unit, respectively. Denoting the latent ability of pupil j in school k by $\nu_{jk}^{(2)}$, we can define the 2-PL model as follows:

$$\text{logit}[P(y_{ijk}=1)] = \beta_i + \lambda_i^{(1)}\nu_{jk}^{(2)} \qquad \text{for } i = 1, ..., I; \tag{3}$$

where $\lambda_i^{(1)}$ is the factor loading or discrimination for item i and $-\beta_i/\lambda_i^{(1)}$ is what is usually referred to as the item difficult. For identification purposes, we will typically restrict one $\lambda_i^{(1)}$, say $\lambda_1^{(1)}$, to be equal to 1. The latent ability is assumed to come from a normal distribution with a mean equal to 0 and a free variance. With the restriction $\lambda_i^{(1)} = 1$ for all i, we obtain the Rasch model.

Suppose we wish to take into account the multilevel structure assuming that that schools belong to one of M latent classes or mixture components with different mean abilities and possibly also different item difficulties. This can be formulated as follows:

$$\text{logit}[P(y_{ijk}=1)] = \beta_i + \lambda_{i1}^{(1)}\nu_{jk}^{(2)} + \sum_{m=1}^{M-1}\lambda_{i,m+1}^{(1)}\nu_{km}^{(3)} \qquad \text{for } i = 1, ..., I \tag{4}$$

$$E(\nu_{jk}^{(2)}) = \sum_{m=1}^{M-1}\lambda_m^{(2)}\nu_{km}^{(3)} \tag{5}$$

$$\text{logit}[P(\nu_{km}^{(3)}=1)] = \gamma_m^{(3)} \qquad \text{for } m = 1, ..., M-1; \tag{6}$$

where $\nu_{km}^{(3)}$ represents one of $M-1$ indicator variables taking the value 1 if school k belongs to latent class m and otherwise 0 (with effect coding $\nu_{km}^{(3)}$ equals -1 if school k belongs to class M). The $\lambda_{i,m+1}^{(1)}$ parameters capture differences between school-level classes in item difficulties and $\lambda_{im}^{(2)}$ in average abilities. In the full model we have to impose identifying constraints on the $\lambda_{i,m+1}^{(1)}$ parameters; for example, $\lambda_{1,m+1}^{(1)} = 0$ for $m = 1, ..., M-1$.

The multilevel mixture IRT model described in equations (4)-(6) can be extended in various ways. The most obvious and interesting extension is inclusion of pupil-level covariates in equation (5) for the latent ability and school-level covariates in equation (6) for the school-level class membership.

The following Latent GOLD 4.5 (Vermunt and Magidson, 2007) syntax file defines the model described in equations (4)-(6):

```
variables
   groupid schoolid;
   caseid childid
   dependent y coding=first;
   independent itemnr nominal;
   latent nu2 continuous, nu3 nominal group 3 coding=last;
equations
   y <- 1 | itemnr + (lambda1) nu2 | itemnr + (lambda2) nu3 | itemnr; // equation for y
   nu2 <- nu3;      // equation for nu2
   nu3 <- 1;        // equation for nu3
```

```
nu2;              // variance of nu2
lambda1[1] = 1;   // discrimination of first item fixed to 1
lambda2[1] = 0;   // item bias of first item fixed to 0
```

This syntax file is rather self explaining: the first part defines the dependent, independent, and latent variables which are in the model, as well as the id variables indicating the three-level data structure. The second part contains the regression equations which are rather similar to equation (4)-(6). The term "`| itemnr`" indicates that a separate constant β_i – denoted by "`1`" – and a separate item discrimination $\lambda_{i1}^{(1)}$ should be estimated for each item. The equations section also contains the specification for the variance of the latent ability and the identifying fixed-value constraint on the discrimination parameter (loading) and item bias of the first item.

Table 2. BIC values obtained with the estimated multilevel mixture 2-PL and Rasch models (N=2156)

number of	2-PL		Rasch	
classes	without item bias	with item bias	without item bias	with item bias
1	40701	40701	40750	40750
2	40502	40545	40562	40517
3	**40449**	40514	40515	40513
4	40455	40502	40524	**40485**
5	40469	40540	40538	40538

Table 2 reports the fit measures obtained with the estimated 1- to 5-class models. As can be seen, the 2-PL models perform better than their Rasch counterparts, indicating that the Rasch assumption of equal discrimination across items is too strict for this data set. For the 2-PL specification, comparison of the models with and without item bias indicates that there is no evidence for item bias. In this specification the 3-class model without item bias is selected as the best according to the BIC criterion. In the Rasch specification, the 4-class model with item bias is the best model. This application shows that using the too restricted Rasch model may lead to the erroneous conclusion that items function differentially across groups.

REFERENCES

Doolaard, S. (1999). Schools in change or school in chain. Phd. dissertation, University of Twente, The Netherlands.

Fox, J. P., and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika, 66, 269-286.

Palardy, G., and Vermunt, J. K. (2007). Multilevel growth mixture models for classifying group-level observations, *submitted*.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal and structural equation models.* London: Chapman & Hall/CRC.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213-239.

Vermunt, J. K. (2004) An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, 58, 220- 233.

Vermunt, J. K. and Magidson, J. (2007). *LG-syntax user's guide: Manual for Latent GOLD 4.5 syntax module.* Belmont, MA: Statistical Innovations Inc.

Item response theory (IRT) has become a popular methodological framework for modeling response data from assessments in education and health; however, its use is not widespread among psychologists. This paper aims to provide a didactic application of IRT and to highlight some of these advantages for psychological test development. IRT was applied to two scales (a positive and a negative affect scale) of a self-report test. Respondents were 853 university students (57 % women) between the ages of 17 and 35 and who answered the scales. IRT analyses revealed that the positive affect scale has items with moderate discrimination and are measuring respondents below the average score more effectively. Item response theory (IRT) is a set of latent variable techniques especially designed to model the interaction between a subjectâ€™s â€œabilityâ€ and the item level stimuli (diculty, guessing, etc.) (Chalmers, 2012).Â For many years CTT remained the dominant framework used in education despite the development and progress of IRT. Currently IRT is nding widespread application in the engineering of large-scale assessments as well as on a smaller scale in sociological and psychological assessments.Â Classical Test Theory versus Item Response Theory. In comparison to classical test theory (CTT), item response theory (IRT) is considered as the standard, if not preferred, method for conducting psychometric evaluations of new and established measures (Osteen 2010). Recent papers in Multilevel Multidimensional Item Response Theory Modeling. Papers. People. à¸à¸²à¸£à¸—à¸”à¸ªà¸šà¹à¸šà¸šà¸›à¸£à¸±à¸šà¹€à¸«à¸¡à¸²à¸°à¸”à¹‰à¸§à¸¢à¸„à¸à¸¡à¸žà¸´à¸§à¹€à¸•à¸à¸£à¹Œà¹à¸šà¸šà¸žà¸«à¸¸à¸¥à¸¹à¸´à¸•à¸´ (Multidimensional Computerized Adaptive Testing : MCAT).Â Educational policy-makers are placing increasing emphasis on testing. All this energy devoted to standardized educational assessment presents a great opportunity for improving instructional decision-making, if testing programs can provide more. Educational policy-makers are placing increasing emphasis on testing. All this energy devoted to standardized educational assessment presents a great opportunity for improving instructional decision-making, if testing programs can provide instructionally meaningful results quickly. Multilevel item response theory (MLIRT) models are used widely in educational and psychological research. This type of modeling has two or more levels, including an item response theory model as the measurement part and a linear-regression model as the structural part, the aim being to investigate the relation between explanatory variables and latent variables. However, the linear-regression structural model focuses on the relation between explanatory variables and latent variables, which is only from the perspective of the average tendency. When we need to explore the relationship between variables at various locations along the response distribution, quantile regression is more appropriate.