

Cross-Classification Multilevel Logistic Models in Psychometrics

Wim Van den Noortgate; Paul De Boeck; Michel Meulders

Journal of Educational and Behavioral Statistics, Vol. 28, No. 4. (Winter, 2003), pp. 369-386.

Stable URL:
http://links.jstor.org/sici?sici=1076-9986%28200324%2928%3A4%3C369%3ACMLMIP%3E2.0.CO%3B2-D

Journal of Educational and Behavioral Statistics is currently published by American Educational Research Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/aera.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Tue May 29 03:43:05 2007

Journal of Educational and Behavioral Statistics
Winter 2003, Vol. 28, No. 4, pp. 369–386

Cross-Classification Multilevel Logistic Models in Psychometrics

Wim Van den Noortgate
Paul De Boeck
Michel Meulders
Katholieke Universiteit Leuven

In IRT models, responses are explained on the basis of person and item effects. Person effects are usually defined as a random sample from a population distribution. Regular IRT models therefore can be formulated as multilevel models, including a within-person part and a between-person part. In a similar way, the effects of the items can be studied as random parameters, yielding multilevel models with a within-item part and a between-item part. The combination of a multilevel model with random person effects and one with random item effects leads to a cross-classification multilevel model, which can be of interest for IRT applications. The use of cross-classification multilevel logistic models will be illustrated with an educational measurement application.

Keywords: *crossed random effects, item response theory, logistic mixed models, multilevel models*

Suppose that a set of items is presented to a group of persons and that for each person the correctness of the responses is recorded. These responses will usually vary, partly at random, partly in a systematic way because of differences in person ability and in item difficulty. An item response theory (IRT) defines the probability of responses of persons to items as a function of person and item characteristics. It has been shown earlier that regular IRT models can be formulated as multilevel logistic models (Adams, Wilson, & Wu, 1997; Kamata, 2001). Making use of this reformulation, we will present a type of IRT model that is based on the principle of cross-classification multilevel models. We will first explain how this principle can be of interest for IRT applications, and we will then discuss the estimation of the unknown parameters of these models. We continue with a discussion of a range of cross-classification logistic models, some of which will be illustrated using an example in which for a group of pupils the attainment targets for reading comprehension are evaluated. We end with a discussion and some conclusions. Although in the following we will consider only dichotomous responses (e.g., correct/incorrect), also categorical responses (e.g., yes/no/perhaps) can be modeled in the way we present.

A Simple Cross-classification Logistic Model

In the basic IRT model, the Rasch model (Rasch, 1960), the response of person j to an item i is regarded as a function of the person ability (θ_j) and the item difficulty (δ_i):

$$\text{Logit}(\pi_{ij}|\theta_j) = \ln(\pi_{ij}/1 - \pi_{ij}) = \theta_j - \delta_i \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}) \quad (1)$$

with $i = 1, \ldots, I$ indicating the item,
$j = 1, \ldots, J$ indicating the person, and
π_{ij} the probability that person j will answer item i correctly.
A common assumption is to regard the persons as a random sample from a population in which the person abilities are normally distributed:

$$\text{Logit}(\pi_{ij}|\theta_j) = \theta_j - \delta_i \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}) \text{ with } \theta_j \sim N(0, \tau^2) \quad (2)$$

Person abilities are regarded exchangeable or otherwise stated it is assumed that permutations of response patterns over persons are equally likely. The only parameters that remain for the persons are the mean and the variance of the person distribution. The parameters can be estimated using a maximum likelihood procedure, often called the marginal maximum likelihood procedure (MML) (Bock & Aitkin, 1981) because the ability parameters are integrated out. Individual person ability parameters can be estimated afterwards, for example using empirical Bayes techniques.

A reformulation of Equation 2 illustrates that the MML formulation of the Rasch model can be regarded as a hierarchical two-level logistic model (Kamata, 2001). The model is a repeated measurement model with responses nested within persons (Figure 1).

The first-level model includes dummy variables, one for each item, as characteristics of the responses. The ith item dummy variable equals 1 if a score is obtained for item i, 0 if the score is obtained for another item (see item dummy variables in Table 1):

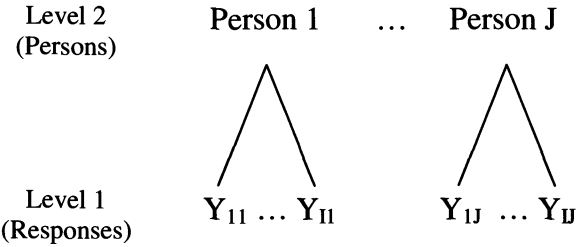


FIGURE 1. *The modeled data hierarchy for a multilevel model with random person effects.*

$$\text{Logit}(\pi_{ij}) = \beta_{0j} + \sum_{k=1}^{I} \beta_k X_{ki} \tag{3}$$

with $Y_{ij} \sim$ Bernoulli(π_{ij}), and
$X_{ki} = 1$ if $k = i$, $X_{ki} = 0$ otherwise

The index j of the intercept indicates that the intercept varies according to the Level-2 units, the persons. At the second level, the person level, we have:

$$\beta_{0j} = u_j \tag{4}$$

with $u_j \sim N(0,\sigma_u^2)$,

Note that in order to make the model identified, the mean of the intercept is constrained to be zero. Combining Equations 3 and 4 gives:

$$\text{Logit}(\pi_{ij}) = \sum_{k=1}^{I} \beta_k X_{ki} + u_j \tag{5}$$

with $Y_{ij} \sim$ Bernoulli(π_{ij}), and $u_j \sim N(0,\sigma_u^2)$
Equations 2 and 5 are equivalent. The coefficients of the item dummy variables in the multilevel logistic model (Equation 5), the βs, correspond to the item difficulty parameters of Equation 2 and are regarded as fixed parameters. The random person effects in the multilevel model, the us, correspond to the random person ability parameters of Equation 2, the variance parameter σ_u^2 to the variance parameter τ^2.

If the purpose of the analysis is to evaluate the abilities of some specific persons, rather than to evaluate the difficulties of some specific items, the choice to regard the person effects as random and the item effects as fixed may seem odd. This may for instance be the case if, to evaluate specific persons, items are generated from a large item bank, possibly based on a set of item design variables. In this case it seems natural, although very uncommon in IRT-analyses, to consider persons as fixed, and items as random. Again, we can use a hierarchical two-level logistic model, to model the responses (Level 1) nested within the Level-2 units, now being the items (Figure 2). This time, person dummy variables, Z_{kj} (see person dummy variables in Table 1), are used as Level-1 covariates (Equation 6). The number of dummy variables equals J, the number of persons:

$$\text{Logit}(\pi_{ij}) = \sum_{k=1}^{J} \beta_k Z_{kj} + u_i \tag{6}$$

with $u_i \sim N(0,\sigma_u^2)$ and $Z_{kj} = 1$ if $k = j$, zero otherwise.
Finally, if both items and persons are regarded as random samples from a population of items and a population of persons respectively, one can define a random residual for both items and persons, with responses nested within persons and within items (see Figure 3).

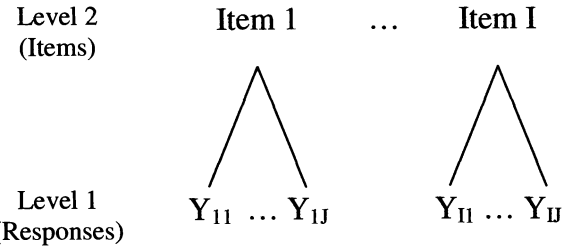


FIGURE 2. *The modeled data hierarchy for a multilevel model with random item effects.*

This multilevel IRT model reads as follows:

$$\text{Logit}(\pi_{ij}) = \beta_0 + u_{1j} + u_{2i} \tag{7}$$

with $u_{1j} \sim N(0,\sigma^2_{u1})$ and $u_{2i} \sim N(0,\sigma^2_{u2})$.

The model does not include dummy variables, but instead we have two Level-2 residual terms. For the estimation, however, dummy variables on the first level may be required, as explained in the section on Estimation. For clarity, we added indices 1 and 2 to refer to the person and item residuals respectively. The model is not a strictly hierarchical model, since items are not nested within persons (every item is offered to all persons), and persons are not nested within items (every person responds to all items). Instead, the item classification and the person classification are found on the same level of the hierarchy and they are crossed in the design. Responses are nested within pairs resulting from a crossing of persons and items. In the multilevel literature, a model with crossed factors is referred to as a cross-classification model or crossed random-effects model (Goldstein, 1987; Raudenbush, 1993). Note that in the special case of IRT models there is mostly only one observation for each cell in the crossed design (i.e., for each pair of a person and an item). For a representation of the cross-classification structure for data in an IRT context, see Table 1. Cross-classification models certainly allow for multiple observations in each cell, but this situation is uncommon in educational measurement applications.

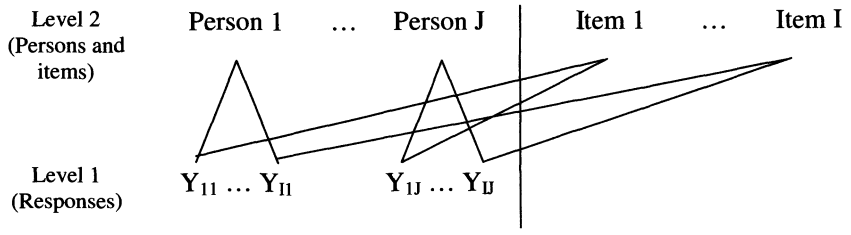


FIGURE 3. *The modeled data structure for a cross-classified 2-level model.*

TABLE 1
Data Structure for a Multilevel Logistic Model Analysis

Response Variable	Item Dummy Variable				Person Dummy Variable			
Y_{ij}	X_{1i}	X_{2i}	...	X_{Ii}	Z_{1j}	Z_{2j}	...	Z_{Jj}
y_{11}	**1**	0	...	0	**1**	0	...	0
y_{21}	0	**1**	...	0	**1**	0	...	0
⋮			...		⋮			
y_{I1}	0	0	...	**1**	**1**	0	...	0
y_{12}	**1**	0	...	0	0	**1**	...	0
y_{22}	0	**1**	...	0	0	**1**	...	0
⋮			...			⋮		
y_{I1}	0	0	...	**1**	0	**1**	...	0
⋮			...					
y_{1J}	**1**	0	...	0	0	0	...	**1**
y_{2J}	0	**1**	...	0	0	0	...	**1**
⋮			...					⋮
y_{IJ}	0	0	...	**1**	0	0	...	**1**

Since the mean of both residual terms from Equation 7 is zero, the intercept equals the estimated logit for the probability of a correct response of an 'average' person on an 'average' item. In the IRT terminology, the intercept thus can be interpreted as the difference between the overall ability and the overall difficulty. We agree that Equation 7 is of less use to the data-analyst, since it explains neither the person abilities nor the item difficulties, and as such it may be considered as a merely descriptive model, without any explanatory value. The model however opens up the perspective of an error term for the items when the issue is to explain the item difficulties from item features (as in Janssen, Tuerlinckx, Meulders, & De Boeck, 2000). The well known Linear Logistic Test Model (LLTM) (Fisher, 1973, 1983) is a model for item difficulties, but it is highly ambitious as it requires perfect predictors since no error term on the item side is included. Equation 7 defines an 'empty' model, but, as will be explained later, when it is complemented with item predictors and possibly with person predictors, it can yield a flexible predictive and explanatory approach, as it includes error terms at both sides.

Estimation

Goldstein (1987) and Raudenbush (1993) demonstrated how cross-classification multilevel models can be formulated as hierarchical multilevel models. A cross-classification model with two types of Level-2 units is reformulated as a model with one type of units as Level-2 units, and the other type as characteristics of the level-1 units, as in Equations 5 and 6. The difference with Equations 5 and 6 is that the coefficients of the response characteristics are now defined to be random effects as well. For example, the Level-2 units could be persons, and the items

could be response characteristics, as in Equation 5, but now with random effects. The way this is done is by defining the coefficients of the item dummy variables to be random on a third (pseudo-)level with only one unit, encompassing the whole data set (see Figure 4).

In Equation 8, we consider persons to be the Level-2 units, and items to be the response characteristics at Level 1:

$$\text{Logit}(\pi_{ij}) = \beta_0 + \sum_{k=1}^{I} u_{2k} X_{ki} + u_{1j} \tag{8}$$

with $(u_{21}, u_{22}, \ldots, u_{2I}) \sim N([0, 0, \ldots, 0], \Omega)$

$u_{1j} \sim N(0, \sigma^2_{u1})$.

Each coefficient of the item dummy variables has its own variance, but all these variances are constrained to be equal, and the covariances are constrained to be zero. The covariance matrix Ω thus is a diagonal matrix with identical elements on the diagonal and all zeroes off the diagonal:

$$\Omega = \boldsymbol{I}_{I \times I} \cdot \sigma^2_{u2}, \tag{9}$$

with $\boldsymbol{I}_{I \times I}$ the identity matrix.

By reformulating cross-classification multilevel models as hierarchical multilevel models, specialized software for hierarchical multilevel models, for instance VARCL (Longford, 1988), HLM (Bryk, Raudenbush, & Congdon, 1996) and MLwiN (Goldstein et al., 1998) can be used to estimate the parameters. In these programs, parameters of multilevel logistic models are usually estimated using approximate quasi-likelihood estimation procedures, using a binomial distribution assumption to define the Level-1 variation. Estimates are based on estimated values

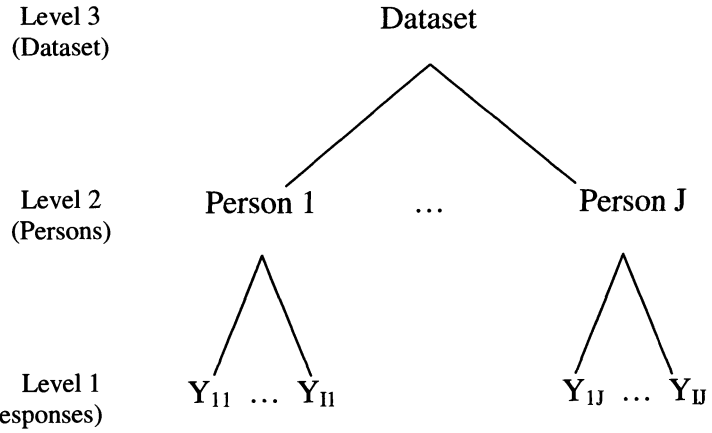


FIGURE 4. *Reformulating the cross-classified 2-level model as a hierarchical 3-level model.*

of the mean and variance of this binomial distribution, which are updated in each iteration. The procedures are known as 'quasi-likelihood' because only the mean and variance of the binomial distribution are used to carry out the estimation (Goldstein, 1995). In the quasi-likelihood procedures, a Taylor expansion is used to approximate and linearize the inverse link function. In the Marginal Quasi-Likelihood procedure (MQL), only the fixed part is used for the Taylor expansion, while in the Penalized or Predictive Quasi-Likelihood procedure (PQL) the Level-2 residual estimates are added to the fixed part when forming the Taylor expansion (Breslow & Clayton, 1993; Goldstein & Rasbash, 1996). In general, the PQL procedure reduces the bias of both fixed and random parameters and therefore is to be preferred (Rodriguez & Goldman, 1995; Goldstein & Rasbash, 1996).

A drawback of reformulating the cross-classification multilevel model as a hierarchical multilevel model is that the covariance matrix on the third level grows quadratically with an increasing number of units (for example, items) that are crossed with the Level-2 units. This may lead to computational problems, making the standard multilevel software practically limited to relatively small problems. To reduce the computational demands, it is preferable to use dummy variables (Level-1 characteristics) for the classification with the smaller number of units (Goldstein, 1987; Raudenbush, 1993), which in IRT-analyses is usually the item classification. The number of classes in the other kind of classification is of much less importance.

Alternatively, the parameters of cross-classification multilevel logistic models can be estimated using SAS. Although the procedure NLMIXED has been developed to estimate the parameters of nonlinear mixed models (e.g., multilevel logistic models), it cannot be used to estimate the parameters of the cross-classification model presented above, because in the procedure only one kind of random effects can be defined, e.g. random person effects or random item effects. Instead, one can use the SAS-macro GLIMMIX (Wolfinger & O'Connell, 1993), in which both MQL and PQL are implemented for fitting generalized linear mixed models. In each iteration, the macro uses a Taylor expansion to linearize the inverse logit function and calls the MIXED procedure for linear mixed models to estimate the parameters.

A third alternative is the use of Bayesian methods. Bayesian methods became more applicable thanks to iterative 'Markov chain Monte Carlo' procedures like the Gibbs sampler or the Metropolis algorithm (Geman & Geman, 1984; Gelman, Carlin, Stern, & Rubin, 1995), and thanks to the development of specialized user-friendly software for Gibbs sampling, such as BUGS (Spiegelhalter, Thomas, Best, & Gilks, 1996). In the Bayesian approach, one is interested in the posterior distribution of the unknown parameters, β_0, σ^2_{u1} and σ^2_{u2} in the case of the multilevel cross-classification IRT model (Equation 7). This posterior distribution, given the observed data and possibly given prior information about the parameters, reflects the uncertainty about the parameter estimates. In the Gibbs sampler, the posterior distribution is not derived analytically, but rather approximated by drawing a sample from the distribution. For that purpose, the posterior distribution of each unknown parameter, conditional on the other parameters, is calculated and samples are drawn successively from these conditional distributions. It can be shown that

once this Markov chain has converged, one samples from the (marginal) posterior distribution of the unknown parameters. An advantage of the Bayesian approach is that the uncertainty of the parameter estimates is entirely taken into account. Moreover, computation of a sample of the entire posterior distribution is advantageous because it provides not only point estimates (i.e., posterior means) of the unknown parameters, but also (1-α)%-posterior intervals without relying on normal approximation of the posterior distribution. As a result, standard errors will also be accurate for small samples (Tanner & Wong, 1987). One of the problems of the Bayesian approach is the evaluation of convergence. For the Gibbs sampler, it is known that, under some mild regularity conditions, the simulated consequences converge to the true posterior distribution (Gelfand & Smith, 1990), but assessing whether convergence has been attained is a difficult problem which has not completely been resolved yet (Cowles & Carlin, 1996). Since the theory underlying the Bayesian approach is relatively complex and the approach is not well known by a large group of researchers, we will focus in the remainder of the manuscript on the (quasi-) maximum likelihood estimation of the parameters of the cross-classification models. For details about the Bayesian approach, we refer to Gelman et al. (1995); for an application to a cross-classification problem, see Janssen, et al. (2000).

Extending the Simple Cross-classification Model

We will present three extensions of Equation 7. First, we discuss the inclusion of a discrimination and a guessing parameter in the models we discussed before, leading to the 2-PL and 3-PL models with random item and/or random person effects. Secondly, external covariates will be included to explain the response probabilities. Finally, the two-level model will be extended by including higher levels. The units on these higher levels may be groups of persons (e.g., schools) and/or groups of items (e.g., specific contents the items refer to).

The 2-PL and 3-PL Model

In some situations, the assumptions underlying the simple Rasch model may be unrealistic. The model assumes, for instance, that all the items discriminate in the same degree between able and less able persons. Birnbaum (1968) proposed an extension of the Rasch model, the so called two parameter logistic (2-PL) model, which includes a second item parameter (α_i) that can be interpreted as an item discrimination parameter:

$$\text{Logit}(\pi_{ij}|\theta_j) = \alpha_i\theta_j - \beta_i = \alpha_i\theta_j - \beta_i. \qquad (10)$$

As is the case for the simple Rasch model, the item parameters of the 2-PL model are usually regarded as fixed parameters, while the person parameter is regarded as random. One could however also regard the items as random and the persons as fixed, or consider both items and persons to be random. The last case leads to a cross-classification 2-PL model with three random parameters. Following the notation we introduced above, the 2-PL model thus reads as follows:

$$\text{Logit}(\pi_{ij}) = \beta_0 + u_{1j}u_{3i} + u_{2i}. \tag{11}$$

The terms on the right-hand side of the 2-PL model (Equations 10 and 11) thus include a multiplication of two unknown parameters: $\alpha_i\theta_j$ or $u_{1j}u_{3i}$. This term can be interpreted as the item-specific effect of the person ability on the probability of a success, regarding the person ability as a latent person covariate. Note that, to ensure that the model is identified, the variance of the person ability parameter (or of the item discrimination parameter) must be constrained, e.g., to 1. If the item parameters are regarded as fixed, one can use the GLIMMIX macro or standard specialized multilevel software (e.g., MLwiN and HLM) to estimate the parameters by alternating the estimation of the person abilities (the θ_j) and the item discrimination parameters (the α_i) (Woodhouse, 1991). However, this iterative procedure is not a standard routine and must be programmed by the user. An analogous procedure can be used for the cross-classification 2PL-model. In this case, the effect of the person ability is assumed to vary randomly over items. When using the GLIMMIX macro or standard multilevel software, the distribution of these random parameters is assumed to be normal. Note that negative discrimination parameters are unlikely and therefore the assumption of normally distributed discrimination parameters is unrealistic. A useful alternative which has been used in a Bayesian framework is to specify a lognormal prior, which constraints the posterior values to be positive (Patz & Junker, 1999). Hence, further adaptations of the program are required, but this topic is beyond the scope of this article.

The Rasch model and the 2-PL model further fail to take into account that, e.g., for a multiple choice item, there may be a non-ignorable chance that a correct answer is due to guessing. Birnbaum (1968) therefore proposed the 3-PL model, which includes, besides the item difficulty parameter and the item discrimination parameter, a third item parameter, that sometimes can be interpreted as a guessing parameter (γ_i):

$$\pi_{ij} = \gamma_i + (1 - \gamma_i)\frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]}. \tag{12}$$

Again one could consider items, persons, or both to be random, and one could formulate the model as a hierarchical or a cross-classification multilevel model, now with three random item parameters. Note that although the model is usually called the three parameter logistic (3-PL) model, it can in fact not be written as a logistic function, as we did for the preceding models. While in the previous models, the logit link function of the expected value of the dependent variable is a linear equation, this is not the case for the 3-PL model. Therefore, the GLIMMIX macro as well as the standard multilevel software, which are designed for hierarchical and non-hierarchical linear and generalized linear models, cannot be used to estimate the parameters of the 3-PL model. Note that when the person effects are fixed, it is possible to treat the item parameters as random when using the procedure NLMIXED from SAS. The parameters of the cross-classification 3-PL model still can be estimated using more complex Bayesian methods.

External Covariates

In Equation 7, item difficulties as well as person abilities are assumed to be independently and identically normally distributed. This means that there is no a priori reason to assume that an item is more difficult than another item and that a specific person would perform better than another person. These assumptions are often unlikely. First, systematic ability effects may be expected of person variables. These person variables may be continuous (e.g., age) or categorical (e.g., gender). Similarly, item characteristics (e.g., number of subtasks or type of problem) can often be assumed to influence the item difficulty. Finally, it is also possible that a person-by-item characteristic has an effect. For example, differential item functioning (DIF) (Holland & Wainer, 1993) can be modeled in this way.

Equation 7 can easily be extended by including characteristics of the three kinds as covariates. The inclusion of covariates can be based on prior beliefs of the researcher about effects of these characteristics, but may also be a tool to explore possible relations. In Equation 13, covariates of these three kinds are included:

$$\text{Logit}(\pi_{ij}) = \beta_0 + \sum_{a=1}^{A} \beta_a^{(1)} M_{ai} + \sum_{b=1}^{B} \beta_a^{(2)} N_{bj} + \sum_{c=1}^{C} \beta_c^{(3)} P_{cij} + u_{1j} + u_{2i}, \qquad (13)$$

with M_a an item covariate, N_b a person covariate, P_c a person-by-item covariate, $\beta_a^{(1)}$, $\beta_b^{(2)}$, and $\beta_c^{(3)}$ the effects of an item, person and person-by-item covariate, and $u_{1j} \sim N(0, \sigma^2_{u1})$ and $u_{2i} \sim N(0, \sigma^2_{u2})$.

Various IRT models that have been proposed before are specific instantiations of Equation 13. The popular Linear Logistic Test Model (LLTM) includes item covariates with fixed effects while person abilities are considered to be random. Since the MML formulation of the Rasch model is a LLTM with dummy item covariates, this model is also a special case of Equation 13. The Random-Effects Linear Logistic Test Model (RE-LLTM) (Janssen, 2002) includes item covariates with fixed effects, but both person abilities and item difficulties (as far as these are not predicted by the item covariates) are assumed to be random. Latent regression IRT models (Zwinderman, 1991) include person covariates (with fixed effects) to model the person abilities. Learning parameters included in dynamic IRT models (Verhelst & Glas, 1993; Verguts & De Boeck, 2000) are effects of covariates of the person-by-item kind. Finally, we note that Equation 13 can be further extended by defining the coefficients of certain item covariates as randomly varying over persons, in the same way as the ordinary LLTM is extended to the Random Weight Linear Logistic Test Model (RWLLTM) (Rijmen & De Boeck, 2002). Similarly, the coefficients of certain person covariates can be defined as random over items.

Additional Levels

A second extension is the definition of higher levels. If the persons can be grouped according to a categorical variable, and the number of groups is small or the groups are the only groups of interest (e.g., male vs. female), the categorical

variable can be included in the model as a person covariate (possibly using dummy variables) with a fixed effect (Snijders & Bosker, 1999). Otherwise, for example when the categorical variable is the school that is attended by the pupils, a new and higher level can be defined. In IRT applications, the same set of items is usually offered in all schools, so that items are not nested within schools and schools are not nested within items. Instead, items are crossed with schools on the third level. If a classification is crossed with units on a specific level, it is also automatically crossed with lower-level units nested within those units (Goldstein, 1995). As a result, the cross-classification of items and schools implies the cross-classification of items and pupils, because the latter are nested within schools. To formulate the model as a hierarchical multilevel model, we define a first level of responses (with items as characteristics), a second level of pupils, a third level of schools, and finally a fourth (pseudo-)level to model the cross-classification of schools and items (also implying the crossing of pupils and items). The model becomes even more complex if also the items can be grouped, for example according to the attainment targets that are evaluated by the items (Figure 5). To reformulate this model as a hierarchical multilevel model, we would need an additional level with one unit encompassing the whole data set, resulting in a hierarchical 5-level model (Figure 6). The coefficients of the dummy variables that indicate the groups of items vary randomly on this fifth level, with variances that are constrained to be equal for all groups. In a similar way as for the persons and items, also characteristics of the schools or the attainment targets can be incorporated as covariates.

Example

The Flemish Community in Belgium issued a set of attainment targets that specify the basic competencies that are expected from pupils leaving primary education. De Boeck, Daems, Meulders, & Rymenans (1997) explored the assessment of the attainment targets for reading comprehension in Dutch. These attainment targets are characterized by the type of text and by the level of processing. In the example, we use the data of one of the tests that were developed by the authors and were reanalyzed by Janssen et al. (2000) as an application of the Random-Effects Linear

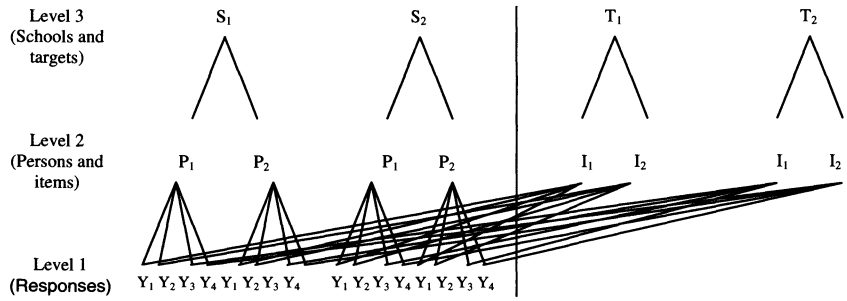


FIGURE 5. *The modeled data structure for a cross-classified 3-level model.*

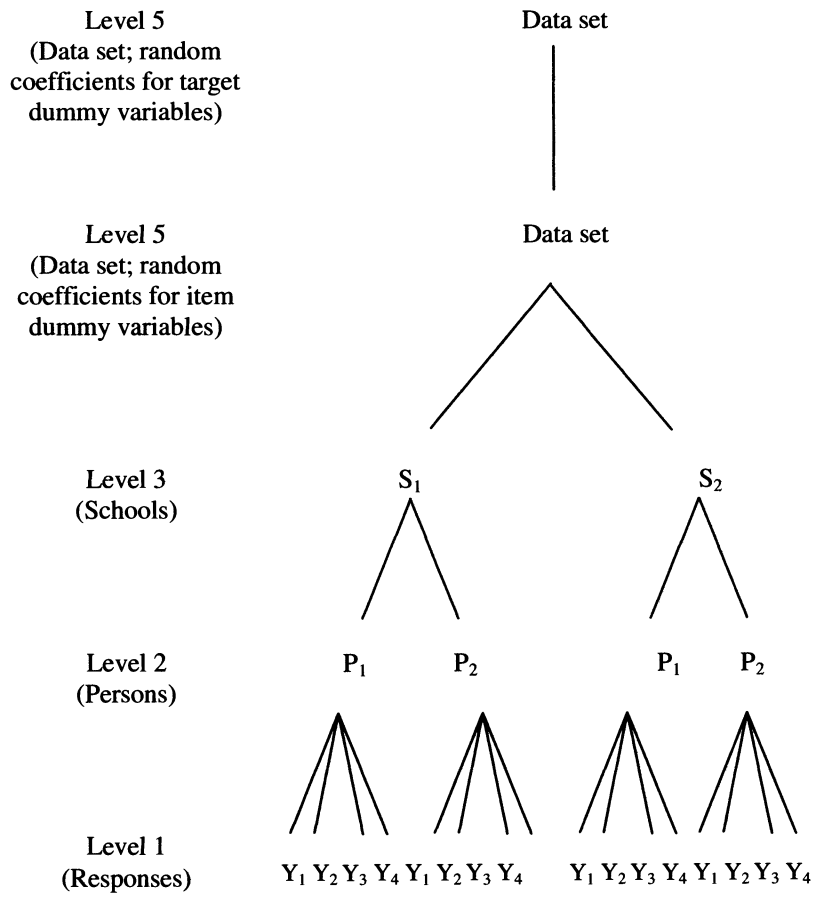


FIGURE 6. *Reformulating the cross-classified 3-level model as a hierarchical 5-level model.*

Logistic Test Model. The data consist of the responses of a group of 539 pupils from 15 schools on 57 items intended to measure 9 attainment targets. In Table 2, the 9 attainment targets that are supposed to be measured by the test are defined by a combination of type of text and level of processing. In addition, we indicate the number of items that measure each target.

We performed multilevel analyses, using the cross-classified logistic multilevel models that were presented above. To estimate the unknown parameters, we used the GLIMMIX macro from SAS, as well as the MLwiN software for hierarchical multilevel models. MQL estimates from both programs are similar. Since MQL estimates in general are shown to be inferior to PQL estimates (Rodriguez & Goldman, 1995), we will focus on the PQL estimates of the parameters. Because MLwiN sometimes crashed when using the PQL estimation procedure, we only

TABLE 2
Text Type, Level of Processing, and Number of Items (I_k) for the Attainment Targets ($k = 1, \ldots, 9$) of the Test

k	Text Type	Level of Processing	I_k
1	Instructions	Retrieving	4
2	Articles in magazine	Retrieving	6
3	Study material	Structuring	8
4	Tasks in textbook	Structuring	5
5	Comics	Structuring	9
6	Stories, novels	Structuring	6
7	Poems	Structuring	8
8	Newspapers for children, textbooks, encyclopedias	Evaluating	6
9	Advertising material	Evaluating	5

Note. From Janssen, Tuerlinckx, Meulders, & De Boeck (2000), used with the permission of the authors.

present the PQL estimates obtained with SAS. These estimates can be found in Table 3. The model is built stepwise. The initial model is a simple descriptive model, but in each step, additional parameters are added. The significance of the parameters is evaluated using the Wald test, comparing the estimates divided by their standard errors with a standard normal distribution. Parameters that do not appear to be significantly different from zero possibly can be omitted in a next step. Note that we do not present the deviance statistics of the models. Deviances are not very accurate since we estimated the parameters using a quasi-likelihood procedure, in which not the real likelihood is maximized. As an unfortunate result, it is unsafe to use the deviance test, comparing the deviance values of the models in order to choose the best model.

TABLE 3
Estimates of the Parameters of the Cross-Classification Multilevel Logistic Models

Parameter	Model 1	Model 2	Model 3	Model 4
Fixed				
Intercept	0.60 (0.15)	0.57 (0.18)	0.56 (0.18)	0.16 (0.35)
Sex			0.15 (0.07)	0.15 (0.07)
Staying down			–0.55 (0.10)	–0.55 (0.10)
Retrieving				0.67 (0.49)
Structuring				0.45 (0.39)
Random				
Intercept				
Target		0.02 (0.12)	0.02 (0.12)	0
Item	1.25 (0.24)	1.24 (0.26)	1.24 (0.26)	1.26 (0.25)
School		0.09 (0.04)	0.07 (0.04)	0.07 (0.04)
Pupil	0.54 (0.04)	0.46 (0.04)	0.43 (0.03)	0.43 (0.03)

Note. Standard errors are given within parentheses.

We start with the fully descriptive model formulated in Equation 7, with responses nested within items and pupils. As shown in the first column of Table 3, the estimate of the intercept equals 0.60. Taking the 'antilogit' of this value, reveals that the expected probability that a pupil gives a correct answer on an item is .65. We further see that the probability of a correct response seems to vary over persons and especially over items. According to the Wald test, both variance parameters appear to be highly significant ($z = 5.25$ and 13.5 respectively, $p < .001$, one tailed[1]). In order to have an idea of the size of the variance, we calculated the expected probability of a successful answer on an average item, for a pupil with an ability of one standard deviation lower, and for a pupil with an ability of one standard deviation higher than the average ability. These probabilities are .47 and .79 respectively, the antilogits of $(0.60 - \sqrt{0.54})$ and $(0.60 + \sqrt{0.54})$, while the probability of a correct answer on an average item is .65. For an average pupil, the probabilities on a correct answer on an item with a difficulty of one standard deviation lower or one standard deviation higher than the average difficulty are .37 and .85 respectively.

Of course, it is possible that pupils within a school perform more similar than pupils from different schools. This means that schools may differ from one another in the overall performance, or that the differences between pupils may be decomposed in variance within schools, and variance between schools. Similarly, the variance between items may be decomposed in variance within attainment targets, and variance between attainment targets. In the second model (second column of Table 3), we further model the multilevel structure, by taking into account the grouping of pupils and items in respectively schools and attainment targets. Differences between schools appear to be relatively small, compared with differences between pupils within the same school: only 16% [= 0.09/(0.09 + 0.46)] of the variance between pupils seems to be attributable to differences in school. Although small, the between-school variance is significant at a .05 alpha level, $z = 2.25$, $p = .012$.

The situation is different at the item side. Only a very small part (2%) of the differences between items appears to be attributable to the attainment targets. Differences in difficulty between the attainment targets are statistically not significant, $z = 0.17$, $p = .43$. The small and non-significant variance could be an argument to drop the attainment target level from the model. For illustrative purposes however, we continue with the model including the attainment targets as a relevant classification of the items.

In the third model (Table 3), we look for an explanation of the differences between pupils in the probability of giving a correct answer, by including 'sex' and 'staying down' as predictors. While the first person covariate equals 1 for girls, 0 for boys, the second covariate equals 1 for pupils that have stayed down at least one year in primary school, and 0 for the other pupils. The coefficient of sex, which is statistically significant at the .05 alpha level, $z = 2.14$, $p = 0.03$, indicates that female pupils perform slightly better: for female pupils (who did not yet stay down), the probability of a correct response on the average item is .67, while for male pupils,

the corresponding probability is .64. These values are derived as the antilogits of (.56 + .15) and of (.56) respectively. In addition, as could be expected, the probability of giving a correct answer is lower for pupils who have stayed down at least one year. For male pupils for example, the probability of a correct response for pupils who have stayed down at least one year is .50, while it is .64 for male pupils who did not yet stay down, $z = -5.50$, $p < .0001$. Note that including the pupil covariates results in a (small) decrease of differences between pupils and between schools, while differences between items and attainment targets are relatively unaffected.

Earlier we said that the level of processing characterizes the attainment targets. Three levels are distinguished (in order of complexity): retrieving, structuring and evaluating. We expect that items corresponding to an attainment level that demands a higher level of processing are more difficult. In the fourth model (Table 3), we included two dummy variables to indicate the level of processing. While the first dummy equals 1 if the level of processing for the attainment target measured by the item is retrieving, and 0 otherwise, the second dummy variable equals 1 if the level of processing is structuring. The third level of processing, evaluating, thus is used as the reference category. The expectation is confirmed by the results: the logits for items requiring retrieving or structuring are respectively 0.67 and 0.45 higher than for items requiring evaluating. For a male pupil who did not yet stay down for example, this means that the expected probabilities of a correct answer are .70, .65, and .53 respectively for items that require retrieving, structuring and evaluating. Although the estimated probabilities for the three levels differ substantially, these differences are not significant at an .05 alpha level ($z = 1.39$, $p = .17$ for retrieving vs. evaluating and $z = 1.18$, $p = .24$ for structuring vs. evaluating). Estimates of the variance components are not strongly influenced. The estimated residual variance between attainment targets however is now zero. This means that conditional on the pupil and target covariates included in the model (i.e., levels of processing), attainment targets do not differ more than could be expected on the basis of the stochastic nature of the response.

Discussion

Approaching IRT problems with cross-classification multilevel logistic models is appealing for several reasons. First, defining the item effects as (partly) random is an efficient and realistic way to model the variation in item difficulty. In addition, it will be clear from the preceding that the cross-classification multilevel logistic model is a very flexible model that generalizes several common IRT models in psychometrics. This generality allows the researcher to build a model according to her/his research interests and hypotheses and avoids some unrealistic assumptions that are made when using common IRT models. Moreover, although in IRT-analyses often exactly one response is available for every combination of a person and an item, the parameters of the cross-classification models can also be estimated if the data are unbalanced (Rasbash & Goldstein, 1994). This means that data of all pupils and items can be used in the analyses, even when some pupils did not respond to every item. Furthermore, we show that the parameters of cross-classification multilevel

logistic models can be estimated using user friendly specialized multilevel software, or using the GLIMMIX macro of SAS.

Unfortunately, there are practical and theoretical problems associated with estimating the models with standard multilevel software. First, in these programs, approximate quasi-likelihood procedures are implemented, which—even the penalized quasi-likelihood procedure—are supposed to be less accurate than numerical integration methods to obtain the parameter estimates. The use of numerical integration methods for the estimation of the parameters of models with more than one kind of random effects (e.g., of cross-classification models, or of hierarchical models with three or more levels) is not impossible, but unfortunately requires huge work. The NLMIXED procedure of SAS, in which numerical integration is used to estimate the parameters of generalized linear or nonlinear mixed models, therefore does not allow defining more than one kind of random effects. A second problem is the computer-intensive nature of the analyses. Although the quasi-likelihood procedures are much faster than numerical integration, each of the analyses performed for the example took several hours on a Pentium III 1.5 Ghz. For the analyses, we used the GLIMMIX macro from SAS. The problem is even larger for specialized multilevel software, because for these programs, the cross-classification model must be reformulated as a hierarchical model, resulting in a huge covariance matrix as described above. Since MLwiN failed when using the PQL procedure to estimate the parameters of the example with 57 items, and even the MQL procedure took several hours to estimate the parameters, the analyses of even larger data sets (possibly with hundreds of items) are likely to be problematic. Due to the increasing computing power of personal computers however, this problem of the size of the model and the computational demands can be expected to become less serious in the future. A final problem associated with using the multilevel software, as well as the GLIMMIX macro, is that they fail to estimate the parameters of a model including a guessing parameter. Estimation for this model can be done by more complex Bayesian estimation procedures.

Notes

[1]Because negative variance estimates are truncated to zero, the null hypothesis of no variance is tested against the one-sided alternative hypothesis that the variance is larger than zero (Snijders & Bosker, 1999). Therefore, one-sided p values are used for testing the significance of the variance parameters, while two-sided p values are used for testing fixed effects.

References

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. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

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

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9–25.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs.* Chicago, IL: Scientific Software International, Inc.

Cowles, K., & Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91,* 883–904.

De Boeck, P., Daems, F., Meulders, M., & Rymenams, R. (1997). *Onwikkeling van een toets voor de eindtermen begrijpend lezen [Construction of a test for the educational target of reading comprehension].* Leuven/Antwerpen (Belgium): University of Leuven/University of Antwerpen.

Fisher, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fisher, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3–26.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London, UK: Chapman & Hall.

Goldstein, H. (1987). Multilevel covariance components models. *Biometrika, 74,* 430–431.

Goldstein, H. (1995). *Multilevel statistical models.* London, UK: Edward Arnold.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 159,* 505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN.* Multilevel Models Project, University of London.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Janssen, R. (April, 2002). *Estimating a random effects version of the linear logistic test model using SAS.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

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

Longford, N. T. (1988). *VARCL: Software for variance component analysis of data with hierarchically nested random effects (Maximum likelihood).* Princeton, NJ: Educational Testing Service.

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

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19,* 337–350.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, DK: The Danish Institute of Educational Research.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18,* 321–349.

Rijmen, F., & De Boeck, P. (2002). The random weights LLTM. *Applied Psychological Measurement, 26,* 271–285.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 158,* 73–90.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* London, UK: Sage Publications.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS: Bayesian inference using Gibbs sampling* [Version 0.5 (Version II)].

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distribution by data augmentation [with discussion]. *Journal of the American Statistical Association, 82,* 528–550.

Verguts, T., & De Boeck, P. (2000). A Rasch model for learning while solving an intelligence test. *Applied Psychological Measurement, 24,* 51–73.

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58,* 395–415.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation, 48.*

Woodhouse, G. (1991). Multilevel item response models. In R. Prosser, J. Rasbash, & H. Goldstein (Eds.), *Data analysis with ML3* (pp. 19–43). London, UK: Institute of Education.

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

Authors

WIM VAN DEN NOORTGATE is a postdoctoral researcher, Department of Education, Katholieke Universiteit Leuven, Vesaliusstraat 2, B-3000 Leuven, Belgium; Wim.VandenNoortgate@ped.kuleuven.ac.be. His areas of specialization are multilevel analysis, meta-analysis, and item response theory.

PAUL DE BOECK is Professor of Psychology, Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102, B-3000 Leuven, Belgium; paul.deboeck@psy.kuleuven.ac.be. His areas of specialization are psychometrics, tests, personality, and intelligence.

MICHEL MEULDERS is a postdoctoral researcher, Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102, B-3000 Leuven, Belgium; michel.meulders@psy.kuleuven.ac.be. His area of specialization is psychometrics.

Manuscript Received July, 2002
Revision Received April, 2003
Accepted April, 2003