PSYCHOMETRIKA—VOL. 79, NO. 1, 84–104
JANUARY 2014
DOI: 10.1007/S11336-013-9360-2

ADDITIVE MULTILEVEL ITEM STRUCTURE MODELS WITH RANDOM RESIDUALS: ITEM MODELING FOR EXPLANATION AND ITEM GENERATION

SUN-JOO CHO

VANDERBILT UNIVERSITY

PAUL DE BOECK

OHIO STATE UNIVERSITY AND KU LEUVEN

SUSAN EMBRETSON

GEORGIA INSTITUTE OF TECHNOLOGY

SOPHIA RABE-HESKETH

UNIVERSITY OF CALIFORNIA, BERKELEY AND INSTITUTE OF EDUCATION, UNIVERSITY OF LONDON

An additive multilevel item structure (AMIS) model with random residuals is proposed. The model includes multilevel latent regressions of item discrimination and item difficulty parameters on covariates at both item and item category levels with random residuals at both levels. The AMIS model is useful for explanation purposes and also for prediction purposes as in an item generation context. The parameters can be estimated with an alternating imputation posterior algorithm that makes use of adaptive quadrature, and the performance of this algorithm is evaluated in a simulation study.

Key words: alternating imputation posterior with adaptive quadrature, item generation, multilevel model, random item parameters.

1. Introduction

The aim of the current study is to extend the two-parameter logistic (2PL) item response model (Birnbaum, 1968) to a model for a multilevel item structure with item attributes for each level of the item structure. The extended model is an integrated framework, which includes item response models for both explanation and item generation. The model can be applied in item generation studies and in all studies with a multilevel item structure and attributes for the levels in that structure. In addition, an approximate maximum likelihood estimation method is provided to estimate the parameters of the extended model.

The 2PL item response model can be written as

$$\text{logit}\big[\Pr(y_{pi}=1|\theta_p)\big]=\eta_{pi}=a_i\cdot\theta_p-b_i, \tag{1}$$

where p is a person index ($p=1,\ldots,P$), i is an item index ($i=1,\ldots,I$), η_{pi} is the logit of the conditional probability $\Pr(y_{pi}=1|\theta_p)$, θ_p is the normally distributed underlying person trait with a mean of 0 and a variance of 1 for model identification, a_i is a slope or discrimination

Requests for reprints should be sent to Sun-Joo Cho, Vanderbilt University, Nashville, USA. E-mail: sj.cho@vanderbilt.edu

© 2013 The Psychometric Society

parameter, and b_i is a location or difficulty parameter. The 2PL item response model does not take a possibly multilevel item structure into account.

A multilevel item structure means that items are nested in *item categories*, such as item families used in an item generation context (Bejar, 1993). As an example of an item family, think of an inductive test with number series of the type "$x_1, x_2, x_3, x_4, \ldots$?" where the next number has to be filled out. Item families can be generated on the basis of a parent item with a given rule, for example $x_t = x_{t-1} + 1 + t$ ($t = 1, \ldots$), as in $(12, 15, 19, 24, \ldots)$ by varying the number to start from, which could lead to items such as $(34, 37, 41, 46, \ldots?)$, $(26, 29, 33, 38, \ldots?)$, etc. constituting an item family.

In addition, for the items and for the categories they are nested in, *attributes* can be considered to model the responses, *item attributes* and *item category attributes*, respectively. For the number series, an example of an item attribute would be low vs. high numbers for series nested within the same inductive rule, such as $(1, 4, 8, 13, \ldots?)$ vs. $(1123, 1126, 1130, 1135, \ldots?)$, and an example of an item category attribute would be whether the inductive rule requires addition vs. multiplication. While these are simple examples for purely illustrative reasons, more complex examples that are more representative of the item generation context can be found in Embretson (2010) and Gierl and Haladyna (2012).

Item generation based on item models is important for the theoretical foundation of test items and their internal validity (Bejar, 1993; Embretson, 1998; Gierl & Haladyna, 2012) as well as for practical reasons such as producing items algorithmically and, therefore, in a more systematic and less time-consuming way, for example, to build an item bank with predictable item parameters (Bejar, 1993, 2012; Embretson, 2010; Gierl & Haladyna, 2012; Gierl & Lai, 2012; Irvine & Kyllonen, 2002). Possible applications include, for example, spatial reasoning (Embretson & Gorin, 2001), mathematical reasoning (Embretson & Daniel, 2008), abstract reasoning (Embretson, 1998), vocabulary (Scrams, Mislevy, & Sheehan, 2002), reading comprehension (Gorin, 2005), and various types of achievement testing (Gierl, Zhou, & Alves, 2008). An automatic Item GeneratOR (IGOR) has been developed for the latter (Gierl et al., 2008).

The extension of the 2PL item response model to be described consists of three elements: (1) two (or more) item levels (multilevel extension), (2) attribute effects per level: effects of *item attributes* and *item category attributes* (attribute effects extension), and (3) a random term per level (random item and category residuals extension). The resulting model is a multilevel regression model for the items. The individual items constitute level 1, and the item categories constitute level 2, while the attributes are the regressors. Like in all regression models, random residuals are included. The extensions of the 2PL item response model are graphically presented in Figure 1. This figure applies to discriminations as well as to difficulties. The three extensions will be further explained in the following.

Multilevel Extension Item generation starts from items that function as models of which variants are generated through a generation process. Together with the item model, the resulting items form an item category. The item model is also called the "parent" item and the generated items are called the "offspring" items. Together they form an item "family." The item structure is therefore multilevel, with individual items at level 1, nested within item categories at level 2. The term 'multilevel' is used for the two levels of the data on the item side. As shown in Figure 1, different kinds of effects can be employed at each level. A random item effect (RI) and a fixed item attribute effect (FIA) can be used at level 1 and a random category effect (RC) and a fixed category attribute effect (FCA) can be used at level 2. Attribute effects and random residuals at the two levels are the next extensions.

Attribute Effects Extension One way to generate items is to create variants based on item attributes (Embretson, 1998, 1999). On the other hand, the items categories can also differ with

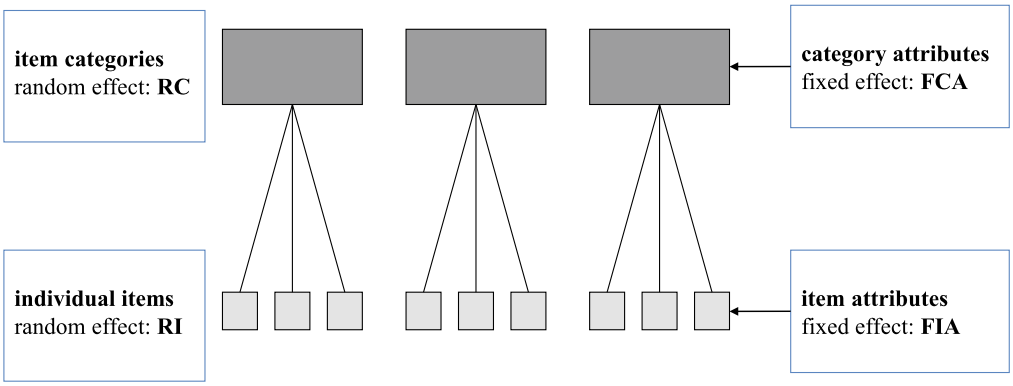


FIGURE 1.
Representation of a multilevel item structure.

respect to their attributes (Geerlings, Glas, & van der Linden, 2011). These attributes of item categories and of items within these categories can have explanatory, and thus also predictive value for the item parameters.

Random Item and Category Residuals Extension In line with the regression model concept, residuals are added for the unexplained variation at both levels. Treating item and category effects as random in this way acknowledges that, apart from the attributes, items and item categories are chosen rather arbitrarily from a broader range of possible choices. Including random category effects is a way to model the dependence between responses to items within the same category that arises from between item-category differences. These reasons for including random category residuals are the same as the reasons for including random intercepts for groups in multilevel regression models for persons nested in groups. We elaborate these ideas further below.

Because an explanation is almost never perfect, an error term should be included, as in regression models, for the item variation that cannot be attributed to the item attributes (Janssen, Schepers, & Perez, 2004), and similarly for the category variation that cannot be attributed to the category attributes. Therefore, random effects (i.e., residuals) at both level 1 and level 2 are needed. It has been shown that omitting the random item residuals can lead to underestimated standard errors for the item attributes effects (Cho, Gilbert, & Goodwin, 2013; Janssen et al., 2004; Mislevy, 1988).

Often, items are sampled from an item bank. One application in educational testing is sampling a different set of items from the same pool of items at different time points to prevent cheating by item exposure (Albers, Does, Imbos, & Janssen, 1989). Like items sampled from an item bank, item categories can be sampled from the item bank because such items function as parents of an item family. Another example is different words sampled from a list of high frequency words (i.e., item category bank) to create items in order to measure lexical representation (Cho et al., 2013). However, it is not necessary to view items or categories as sampled from respective populations. We rather assume exchangeability of the item and category residuals, given the item and category attributes.

A multilevel structure induces dependencies. In the common case of persons nested within groups, the shared group can and mostly does make the observations from different persons within the group more similar than from persons from different groups. In a similar way, the fact that items belong to a common item category can make the items more similar, with consequences for their item parameters. In other words, the multilevel structure of items is a source of dependencies between item parameters. A random category effect accounts for the dependencies.

When the item parameters are predicted by the posterior means, the predictions borrow strength from other items belonging to the same category, and are shrunken toward the category-specific means. In contrast, a fixed effects approach, in which separate fixed parameters a_i and b_i are specified for each item as in Equation (1), does not exploit the dependencies. Glas and van der Linden (2003) and Sinharay, Johnson, and Williamson (2003) have shown that the fixed-effects approach produces greater bias and mean absolute estimation error (Glas & van der Linden, 2003, p. 260), and wider confidence bands for the item parameters (Sinharay et al., 2003, pp. 305–306). As pointed out by Sinharay et al. (2003), the two approaches become indistinguishable as the number of examinees that respond to each item becomes extremely large because the shrinkage then becomes negligible. A further advantage of the random-effects approach is that the item parameters and response functions of newly generated items can be predicted, using the family expected response function (Sinharay et al., 2003) for the latter.

The combination of these three extensions defines a multilevel regression model for the items. The extensions considered here are for both item parameters of the 2PL item response model, the discrimination and the difficulty parameters. Because the attribute effects are modeled as additive for both the difficulties and the discriminations, the model is called the *additive multilevel item structure model* or AMIS model. The full AMIS model is a new model, but as will be discussed in the next section, some of the features of the model have been proposed earlier (Geerlings et al., 2011; Glas & van der Linden, 2003; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000; Johnson & Sinharay, 2005; Sinharay et al., 2003). We develop and apply an approximate maximum likelihood estimation method, as an alternative to the Bayesian approach that is adopted in the earlier item generation studies. The model is applied to a mathematical aptitude test, and the consequences of misspecifying the model by omitting the random item category effects are investigated using a simulation study.

The AMIS model has a range of applications beyond item generation. Each time the item structure is multilevel, with item categories and items nested in these categories, the AMIS model may be considered if discrimination and/or difficulty predictors are available. Two examples are testlets (Wainer, Bradlow, & Wang, 2007) and criterion item sets (Janssen et al., 2000). Testlets are bundles of items with an important element in common, such as the item stem. Criterion item sets are mutually exclusive sets of items measuring a common mastery criterion per set, for example in a criterion-referenced measurement approach. The model is also useful for testing theories about item attributes that may explain item difficulties and discriminations. For example, cognitive complexity can be tested via the coefficients of item attributes in the model for the item difficulties (e.g., Cho, Athay, & Preacher, 2013).

In the first section hereafter, the AMIS model will be described. Brief literature reviews of related models and item generation are provided. In the second section, an approximate maximum likelihood estimation method will be presented to estimate the parameters of the model. An empirical study will be provided in the third section, followed by the simulation study in the fourth section. In the last section, we discuss the model in the light of the empirical study and the simulation study.

2. Additive Multilevel Item Structure Model

The full AMIS model is expressed in the following equation for the logit of person p responding to item i belonging to category c:

$$\begin{aligned}\eta_{pic} = &\left(\mu_\alpha + \sum_d \gamma_{\alpha d} Q_{id} + \varepsilon^{(1)}_{\alpha i} + \sum_t \delta_{\alpha t} R_{ct} + \varepsilon^{(2)}_{\alpha c}\right) \cdot \theta_p \\ &- \left(\mu_\beta + \sum_d \gamma_{\beta d} Q_{id} + \varepsilon^{(1)}_{\beta i} + \sum_t \delta_{\beta t} R_{ct} + \varepsilon^{(2)}_{\beta c}\right). \end{aligned} \tag{2}$$

The equation has the same structure as Equation (1) for the 2PL item response model. It contains a discrimination term, a difficulty term, and a person trait term. The person trait term is not different from the 2PL item response model, but the discrimination and the difficulty terms are each modeled using several components. The discrimination (a_i in Equation (1)) is modeled as the intercept (μ_α) and four other components:

1. $\sum_d \gamma_{\alpha d} Q_{id}$: the effect of the level-1 attributes, where Q_{id} is the value of item attribute d for item i, and $\gamma_{\alpha d}$ is the effect of attribute d. This component is the fixed item attribute (FIA) component of the discrimination.
2. $\varepsilon_{\alpha i}^{(1)}$: the level-1 residual, unexplained by the item attributes, $\varepsilon_{\alpha i}^{(1)} \sim N(0, \psi_\alpha^{(1)})$. This component is the random item (RI) component of the discrimination.
3. $\sum_t \delta_{\alpha t} R_{ct}$: the effect of the level-2 attributes, where R_{ct} is the value of category attribute t for item category c, and $\delta_{\alpha t}$ as the effect of attribute t. This component is the fixed category attribute (FCA) component of the discrimination.
4. $\varepsilon_{\alpha c}^{(2)}$: the level-2 residual, unexplained by the item category attributes, $\varepsilon_{\alpha c}^{(2)} \sim N(0, \psi_\alpha^{(2)})$. This component is the random category (RC) component of the discrimination.

In a similar way, the difficulty (b_i in Equation (1)) is modeled as the intercept (μ_β) and four other components:

1. $\sum_d \gamma_{\beta d} Q_{id}$, where $\gamma_{\beta d}$ is the effect of the level-1 attribute d. This is the FIA component of the difficulty.
2. $\varepsilon_{\beta i}^{(1)}$, the level-1 residual, unexplained by the item attributes, $\varepsilon_{\beta i}^{(1)} \sim N(0, \psi_\beta^{(1)})$. This is the RI component of the difficulty.
3. $\sum_t \delta_{\beta t} R_{ct}$, where $\delta_{\beta t}$ is the effect of the level-2 attribute t. This is the FCA component of the difficulty.
4. $\varepsilon_{\beta c}^{(2)}$, the level-2 residual, unexplained by the item category attributes, $\varepsilon_{\beta c}^{(2)} \sim N(0, \psi_\beta^{(2)})$. This is the RC component of the difficulty.

The model in Equation (2) can be extended by allowing for heteroscedastic residuals and a bivariate distribution for the residuals of discrimination and difficulty. For this article, we have opted for homogeneous distributions and independence. This latter option is more plausible when a larger share of item variance is explained by the item and item category attributes. Another choice concerns the kind of distribution. Existing practices do not differ with respect to the difficulty parameters, where a normal distribution is invariably used. The situation is more complex for the distribution of the discriminations. Mislevy (1986), Glas and van der Linden (2003), Sinharay et al. (2003), and Soares, Gonçalvez, and Gamerman (2009) specify a lognormal distribution, while Bradlow, Wainer, and Wang (1999), Klein Entink, Fox, and van der Linden (2009a), and Geerlings et al. (2011) work with a normal distribution. Given the additive structure for the discrimination parameters, the normal distribution seems a logical choice. In fact, Klein Entink et al. (2009a) and Geerlings et al. (2011) also use a linear function for the discrimination and use the normal distribution. The lognormal distribution for discriminations on the other hand is natural in multiplicative (vs. additive) models for discriminations. Our choice of an additive function is partly inspired by the theoretical possibility that discriminations can be negative, for example, in the personality and attitude domains.

2.1. Related Models

Multilevel Item Models Except for Janssen et al. (2000), all models thus far with item categories and within-category random effects have been described for the purpose of modeling item generation data. They all share a RI component and a second level. A prominent model is the "Related Siblings Model" described by Sinharay et al. (2003), which is also used by Glas and

van der Linden (2003), and by Johnson and Sinharay (2005). The item parameter distributions depend on category-specific hyper-parameters with diffuse hyper-prior distributions. The hyper-parameters that correspond to the category-specific means of the item parameters can be viewed as fixed category effects (FC) instead of the RC component of the full AMIS model. The linear item cloning model (LICM) of Geerlings et al. (2011) does make use of attributes for level 2, but without a random residual, as will be discussed in the second type of related models, Explanatory Models with Residuals, below. One can of course also define a single-level model, with fixed effects for each single item (FI), taking us back to the standard 2PL item response model.

Explanatory Models with Residuals Mostly, item predictors do not provide a perfect explanation. Therefore, Janssen et al. (2004) have formulated an explanatory item model with residuals, although only for the difficulties. Their model is an extension of the linear logistic test model (LLTM, Fischer, 1973). Klein Entink, Kuhn, Hornke, and Fox (2009b) have used a model that also includes predictors and a residual for the discriminations. Much earlier, Embretson (1999) had described a model with predictors for both the item difficulties and the item discriminations, but without residual terms. All these models have a FIA component, and except for Embretson's model also a RI component. Finally, the linear item cloning model (LICM, Geerlings et al., 2011) makes use of category attributes (FCA), while it has no residual at level 2 (no RC) and neither does it make use of item attributes. The within category variation is modeled exclusively with random item effects (RI). The category attributes are used in the LICM only for the item difficulties. Therefore, the model is a mixed FCA&RI and FC&RI model.

Random Item Models Traditional item response models do not include *random item effects* (random across items), as in Equation (2). However, there is a growing number of publications with random item effect models (Chaimongkol, Huffer, & Kamata, 2006; De Jong & Steenkamp, 2010; De Jong, Steenkamp, & Fox, 2007; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010, 2011; Glas & van der Linden, 2003; Janssen et al., 2000; Johnson & Sinharay, 2005; Klein Entink et al., 2009a, 2009b; Soares et al., 2009; Sinharay et al., 2003; van der Linden, Klein Entink, & Fox, 2010). All these models contain a RI component. A discussion of random item models can be found in De Boeck (2008).

2.2. *Item Generation*

Item generation is becoming important for building an item bank, for adaptive testing, or even to generate items on the fly during adaptive testing. Two kinds of approaches are described in the literature: the item cloning approach and the item attribute approach.

Item Cloning Approach The approach consists of reproducing items through cloning. Glas and van der Linden (2003) describe a two-staged item cloning procedure in which existing items are selected as "parent items" and cloned into item family members. An item family can be created, for example, using a syntactic formulation of the parent item with one or more open slots to be filled in from substitution sets to be specified (Millman & Westman, 1989), or by linguistic transformations (Bormuth, 1970). Elements of the parent items are replaced with alternative similar elements (Roid & Haladyna, 1982). Another term for parent item or item template is "item model" (not in the statistical sense), although the latter term also refers to the specifications for authoring/generating items (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003).

Item Attribute Approach The attribute-based approach consists of constructing items from a set of attributes (Freund, Hofer, & Holling, 2008; Holling, Bertling, & Zeuch, 2009) and is nicely in line with the structural modeling approach of Embretson (1998). Irvine and Kyllonen

(2002) have introduced the terminological pair *incidentals* and *radicals*. Incidentals can be considered as surface features, which do not exert an influence on the item parameters while radicals refer to item features that systematically affect item parameters. The incidentals and radicals may be understood as referring to the difference between cloning and attribute variation as a basis for item generation.

Combined Approaches The two approaches, cloning and attribute variation, can be combined in two ways. The first way is to use attributes for the parent items, and thus for the item categories. In other words, it is a cloning approach supplemented with parent attributes, as in the LICM of Geerlings et al. (2011). The second way of combining the two approaches is to generate offspring items from parent items through "genetic manipulation," a metaphor for the manipulated variation of attribute values in the offspring. This combination is described by Embretson and Yang (2007) and Daniel and Embretson (2010). In Daniel and Embretson (2010) existing items play the role of parent items and structural variants of the parent items are created based on cognitive attributes. The model used by Daniel and Embretson (2010) is a FC&FIA model, with fixed item category effects and fixed item attribute effects. Finally, one can have the simultaneous application of the two ways of combining approaches described above: (1) starting from systematically different parent items (differing with respect to category attributes) and (2) creating attribute variation in the offspring. As will be explained, the generation procedure of Daniel and Embretson (2010) can be reinterpreted as such, since the parents do differ systematically on a number of attributes. The most appropriate model for the resulting item structure is a model with two levels and with attribute effects at both levels, which is the AMIS model of Equation (2).

In an item generation context, modeling the items can have two aims. First, given a dataset based on generated items, one may want to estimate the person traits (e.g., abilities) relying on a given statistical item model. Second, one may want to have an idea of the difficulty and discrimination of a generated item before it is generated, for example because one wants it to be optimally informative in an adaptive procedure. With such a purpose in mind, the item model needs to have a reasonable predictive value.

3. Parameter Estimation

The random effects $\varepsilon_{\alpha i}^{(1)}$, $\varepsilon_{\beta i}^{(1)}$, $\varepsilon_{\alpha c}^{(2)}$, and $\varepsilon_{\beta c}^{(2)}$ for items in Equation (2) are crossed with the random effects for persons θ_p. Maximum likelihood estimation of models for categorical data with crossed random effect is challenging. This is because the marginal likelihood does not have a closed form so that estimation requires numerical or Monte Carlo integration. If the random effects are nested, the integrals are also nested (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2005), keeping the computational burden low, but for crossed random effects, high-dimensional integrals need to be evaluated.

Approximations to maximum likelihood estimation have been proposed to avoid high-dimensional numerical integration in generalized linear mixed models, including marginal quasi-likelihood (MQL, Goldstein, 1991), penalized quasi-likelihood (PQL, Breslow & Clayton, 1993), its second-order improvement (PQL-2, Goldstein & Rasbash, 1996), bias-corrected PQL (Breslow & Lin, 1995; Lin & Breslow, 1996), Laplace approximations (Tierney & Kadane, 1986; Pinheiro & Bates, 1995; Raudenbush, Yang, & Yosef, 2000), and the hierarchical-likelihood method (Lee & Nelder, 1996, 2006). However, MQL, PQL, and standard Laplace methods tend to underestimate the variances for dichotomous response with small cluster sizes and large variance components (Breslow, 2004; Browne & Draper, 2006; Cho & Rabe-Hesketh, 2011; Joe, 2008; Rodriguez & Goldman, 1995, 2001).

Markov chain Monte Carlo (MCMC) has been used as a straightforward approach for the estimation of models with crossed random effects (Karim & Zeger, 1992; Rasbash & Browne, 2007). However, MCMC is computationally expensive, and it is difficult to specify vague priors for the variance parameters in hierarchical models that result in a posterior mean (or mode) close to the maximum likelihood estimate (MLE) (Browne & Draper, 2006; Natarajan & Kass, 2000). Furthermore, convergence may be slow if there are strong correlations in the joint posterior distribution of the parameters (Clayton & Rasbash, 1999).

The alternating imputation-posterior algorithm (AIP, Clayton & Rasbash, 1999) makes use of data augmentation so that a sequence of simpler models can be estimated. The aim is to obtain approximate MLEs. Unlike MCMC, the AIP algorithm does not require specification of prior distributions for the model parameters. Furthermore, the algorithm typically converges much more rapidly because several model parameters are updated simultaneously. Clayton and Rasbash (1999) used marginal quasi-likelihood (MQL-2) and penalized quasi-likelihood (PQL-2) to estimate the simpler models within the AIP algorithm. Because these approximations are known to perform poorly in many situations, Cho and Rabe-Hesketh (2011) developed an AIP algorithm that uses maximum likelihood estimation with adaptive quadrature (Bock & Schilling, 1997; Schilling & Bock, 2005; Pinheiro & Bates, 1995; Rabe-Hesketh et al., 2005).

In this study, AIP with adaptive quadrature, as described by Cho and Rabe-Hesketh (2011), is extended to deal with the multilevel item structure of the AMIS model. The full AMIS model having FCA&RC&FIA&RI components was chosen to describe the AIP algorithm with adaptive quadrature below.

3.1. AIP with Adaptive Quadrature

The AIP algorithm is based on the imputation posterior (IP) algorithm of Tanner and Wong (1987) which can be outlined as follows:

I-step (data augmentation): Impute missing data (random effects) sampling from the distribution of the missing data conditional on the observed data. This requires first sampling the parameters from the current approximation of their posterior distribution.

P-step: Update the approximation of the posterior distribution.

For the full AMIS model (Equation (2)), the algorithm consists of three wings, a person wing, an item wing, and an item category wing. The item category wing is a way to deal with the multilevel extension of the model. We let $\boldsymbol{\gamma}$ denote the vectors of regression coefficients for item attributes and $\boldsymbol{\delta}$ denote the vectors of regression coefficients for category attributes. In the person wing, the item and category residuals are treated as known and the parameters, μ_α, μ_β, $\boldsymbol{\gamma}_\alpha$, $\boldsymbol{\gamma}_\beta$, $\boldsymbol{\delta}_\alpha$, and $\boldsymbol{\delta}_\beta$ are estimated (P-step). The person trait value θ_p are sampled (I-step) by first sampling the parameters from their approximate posterior distribution (given the item and category residuals). In the item wing, the person trait values and the category residuals are treated as known and the parameters, μ_α, μ_β, $\boldsymbol{\gamma}_\alpha$, $\boldsymbol{\gamma}_\beta$, $\boldsymbol{\delta}_\alpha$, $\boldsymbol{\delta}_\beta$, $\log \psi_\alpha^{(1)}$, and $\log \psi_\beta^{(1)}$ are estimated (P-step). The item residuals $\varepsilon_{\alpha i}^{(1)}$ and $\varepsilon_{\beta i}^{(1)}$ are sampled (I-step), by first sampling the parameters from their approximate posterior distribution (given the person trait values and category residuals). In the item category wing, the person trait values and item residuals are treated as known and the parameters, μ_α, μ_β, $\boldsymbol{\gamma}_\alpha$, $\boldsymbol{\gamma}_\beta$, $\boldsymbol{\delta}_\alpha$, $\boldsymbol{\delta}_\beta$, $\log \psi_\alpha^{(2)}$, and $\log \psi_\beta^{(2)}$ are estimated (P-step). The category residuals $\varepsilon_{\alpha c}^{(2)}$ and $\varepsilon_{\beta c}^{(2)}$ are sampled (I-step), again by first sampling the parameters from their approximate posterior distribution (given the person trait values and item residuals).

Specifically, after setting initial values $\boldsymbol{\epsilon}_\alpha^{(1)0}$ and $\boldsymbol{\epsilon}_\beta^{(1)0}$ for the item residuals and $\boldsymbol{\epsilon}_\alpha^{(2)0}$ and $\boldsymbol{\epsilon}_\beta^{(2)0}$ for the category residuals, the person wing, item wing, and item category wing outlined below are alternated until convergence. In iteration k:

3.2. Person Wing

Treat the item residuals $\boldsymbol{\epsilon}_{\alpha}^{(1)k-1} = (\varepsilon_{\alpha 1}^{(1)k-1}, \ldots, \varepsilon_{\alpha I}^{(1)k-1})'$, $\boldsymbol{\epsilon}_{\beta}^{(1)k-1} = (\varepsilon_{\beta 1}^{(1)k-1}, \ldots, \varepsilon_{\beta I}^{(1)k-1})'$ and category residuals $\boldsymbol{\epsilon}_{\alpha}^{(2)k-1} = (\varepsilon_{\alpha 1}^{(2)k-1}, \ldots, \varepsilon_{\alpha C}^{(2)k-1})'$, and $\boldsymbol{\epsilon}_{\beta}^{(2)k-1} = (\varepsilon_{\beta 1}^{(2)k-1}, \ldots, \varepsilon_{\beta C}^{(2)k-1})'$, from the previous iteration as known. The model can then be written as

$$\begin{aligned}
&\operatorname{logit}\left[\Pr\left(y_{pic}=1\middle|\theta_p, \varepsilon_{\alpha i}^{(1)k-1}, \varepsilon_{\beta i}^{(1)k-1}, \varepsilon_{\alpha c}^{(2)k-1}, \varepsilon_{\beta c}^{(2)k-1}\right)\right] \\
&\quad = -\left(\mu_\beta + \sum_d \gamma_{\beta d} Q_{id} + \sum_t \delta_{\beta t} R_{ct} + \varepsilon_{\beta i}^{(1)k-1} + \varepsilon_{\beta c}^{(2)k-1}\right) \\
&\qquad + \theta_p\left(\mu_\alpha + \sum_d \gamma_{\alpha d} Q_{id} + \sum_t \delta_{\alpha t} R_{ct} + \varepsilon_{\alpha i}^{(1)k-1} + \varepsilon_{\alpha c}^{(2)k-1}\right),
\end{aligned} \tag{3}$$

where the first term in parentheses is a standard linear predictor with $(\varepsilon_{\beta i}^{(1)k-1} + \varepsilon_{\beta c}^{(2)k-1})$ treated as an offset (covariate with coefficient set to 1) and the term multiplied by θ_p is another linear predictor, with $(\varepsilon_{\alpha i}^{(1)k-1} + \varepsilon_{\alpha c}^{(2)k-1})$ as an offset. The response model of generalized linear latent and mixed models (GLLAMMs, Rabe-Hesketh, Skrondal, & Pickles, 2004) allows each latent variable to be multiplied by a linear predictor, and the above model can therefore be estimated using the **gllamm** command (Rabe-Hesketh et al., 2004, 2005) in Stata. Let the parameters for the person wing be denoted $\boldsymbol{\vartheta}_1 = \{\mu_\alpha, \mu_\beta, \boldsymbol{\gamma}_\alpha, \boldsymbol{\gamma}_\beta, \boldsymbol{\delta}_\alpha, \boldsymbol{\delta}_\beta\}$.

1. Obtain MLEs $\widehat{\boldsymbol{\vartheta}}_1^k$ with estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\vartheta_1}^k$.
2. Sample parameters $\boldsymbol{\vartheta}_1^k$ from their approximate posterior distribution

$$\boldsymbol{\vartheta}_1^k \left|\boldsymbol{\epsilon}_{\alpha}^{(1)k-1}, \boldsymbol{\epsilon}_{\beta}^{(1)k-1}, \boldsymbol{\epsilon}_{\alpha}^{(2)k-1}, \boldsymbol{\epsilon}_{\beta}^{(2)k-1} \right. \sim N\left(\widehat{\boldsymbol{\vartheta}}_1^k, \widehat{\boldsymbol{\Sigma}}_{\vartheta_1}^k\right). \tag{4}$$

3. Sample $\boldsymbol{\theta}^k$ from its posterior distribution with parameters $\boldsymbol{\vartheta}_1^k$.

3.3. Item Wing

Treat the category residuals $\boldsymbol{\epsilon}_{\alpha}^{(2)k-1}$ and $\boldsymbol{\epsilon}_{\beta}^{(2)k-1}$ from the previous iteration and the person trait values $\boldsymbol{\theta}^k$ from the person wing as known. The model can then be written as

$$\begin{aligned}
&\operatorname{logit}\left[\Pr\left(y_{pic}=1\middle|\varepsilon_{\alpha i}^{(1)}, \varepsilon_{\beta i}^{(1)}, \varepsilon_{\alpha c}^{(2)k-1}, \varepsilon_{\beta c}^{(2)k-1}, \theta_p^k\right)\right] \\
&\quad = \left(\mu_\alpha \theta_p^k + \sum_d \gamma_{\alpha d} Q_{id} \theta_p^k + \sum_t \delta_{\alpha t} R_{ct} \theta_p^k + \varepsilon_{\alpha c}^{(2)k-1} \theta_p^k - \mu_\beta - \sum_d \gamma_{\beta d} Q_{id}\right. \\
&\qquad \left. - \sum_t \delta_{\beta t} R_{ct} - \varepsilon_{\beta c}^{(2)k-1}\right) + \left(-\varepsilon_{\beta i}^{(1)} + \varepsilon_{\alpha i}^{(1)} \theta_p^k\right),
\end{aligned} \tag{5}$$

where the first term in parentheses is a standard linear predictor with interactions among the item attributes and the person trait (and with $\varepsilon_{\alpha c}^{(2)k-1}\theta_p^k$ and $-\varepsilon_{\beta c}^{(2)k-1}$ as offsets) and where the second term in parentheses is the same as the random part of a random-coefficient model with random intercept $\varepsilon_{\beta i}^{(1)}$ and random slope $\varepsilon_{\alpha i}^{(1)}$ of θ_p^k. This model can be estimated using the **xtmelogit** command in Stata (Rabe-Hesketh & Skrondal, 2012, Chapter 10). Let the parameters of the item wing be denoted $\boldsymbol{\vartheta}_2 = \{\mu_\alpha, \mu_\beta, \boldsymbol{\gamma}_\alpha, \boldsymbol{\gamma}_\beta, \boldsymbol{\delta}_\alpha, \boldsymbol{\delta}_\beta, \log \psi_\alpha^{(1)}, \log \psi_\beta^{(1)}\}$.

1. Obtain MLEs $\widehat{\boldsymbol{\vartheta}}_2^k$ with estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\vartheta_2}^k$.

2. Sample parameters $\boldsymbol{\vartheta}_2^k$ from their approximate posterior distribution

$$\boldsymbol{\vartheta}_2^k \big| \boldsymbol{\epsilon}_\alpha^{(2)k-1}, \boldsymbol{\epsilon}_\beta^{(2)k-1}, \boldsymbol{\theta}^k \sim N\big(\widehat{\boldsymbol{\vartheta}}_2^k, \widehat{\boldsymbol{\Sigma}}_{\vartheta_2}^k\big). \tag{6}$$

3. Sample $\boldsymbol{\epsilon}_\alpha^{(1)k}$ and $\boldsymbol{\epsilon}_\beta^{(1)k}$ from their posterior distribution with parameters $\boldsymbol{\vartheta}_2^k$.

3.4. Item Category Wing

Treat the item residuals $\boldsymbol{\epsilon}_\alpha^{(1)k}$ and $\boldsymbol{\epsilon}_\beta^{(1)k}$ from the item wing and the person trait values $\boldsymbol{\theta}^k$ from the person wing as known. The model can then be written as

$$\begin{aligned}
&\text{logit}\big[\Pr\big(y_{pic} = 1 \big| \varepsilon_{\alpha i}^{(1)k}, \varepsilon_{\beta i}^{(1)k}, \varepsilon_{\alpha c}^{(2)}, \varepsilon_{\beta c}^{(2)}, \theta_p^k\big)\big] \\
&= \Bigg(\mu_\alpha \theta_p^k + \sum_d \gamma_{\alpha d} Q_{id} \theta_p^k + \sum_t \delta_{\alpha t} R_{ct} \theta_p^k + \varepsilon_{\alpha i}^{(1)k} \theta_p^k - \mu_\beta - \sum_d \gamma_{\beta d} Q_{id} \\
&\quad - \sum_t \delta_{\beta t} R_{ct} - \varepsilon_{\beta i}^{(1)k}\Bigg) + \big(-\varepsilon_{\beta c}^{(2)} + \varepsilon_{\alpha c}^{(2)} \theta_p^k\big),
\end{aligned} \tag{7}$$

where the first term in parentheses is a standard linear predictor with interactions among the item attributes and the person trait (and with $\varepsilon_{\alpha i}^{(1)k} \theta_p^k$ and $-\varepsilon_{\beta i}^{(1)k}$ as offsets), and where the second term in parentheses is the same as the random part of a random-coefficient model with random intercept $\varepsilon_{\beta c}^{(2)}$ and random slope $\varepsilon_{\alpha c}^{(2)}$ of θ_p^k. This model can be estimated using the **xtmelogit** command in Stata. Let the parameters of the category wing be denoted $\boldsymbol{\vartheta}_3 = \{\mu_\alpha, \mu_\beta, \boldsymbol{\gamma}_\alpha, \boldsymbol{\gamma}_\beta, \boldsymbol{\delta}_\alpha, \boldsymbol{\delta}_\beta, \log \psi_\alpha^{(2)}, \log \psi_\beta^{(2)}\}$.

1. Obtain MLEs $\widehat{\boldsymbol{\vartheta}}_3^k$ with estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\vartheta_3}^k$.
2. Sample parameters $\boldsymbol{\vartheta}_3^k$ from their approximate posterior distribution

$$\boldsymbol{\vartheta}_3^k \big| \boldsymbol{\epsilon}_\alpha^{(1)k}, \boldsymbol{\epsilon}_\beta^{(1)k}, \boldsymbol{\theta}^k \sim N\big(\widehat{\boldsymbol{\vartheta}}_3^k, \widehat{\boldsymbol{\Sigma}}_{\vartheta_3}^k\big). \tag{8}$$

3. Sample $\boldsymbol{\epsilon}_\alpha^{(2)k}$ and $\boldsymbol{\epsilon}_\beta^{(2)k}$ from their posterior distribution with parameters $\boldsymbol{\vartheta}_3^k$.

After convergence is achieved, the algorithm is continued for a fixed number of iterations and the parameter estimates are obtained by averaging the estimates obtained after burn-in. We refer to Cho and Rabe-Hesketh (2011) for details on convergence checking and posterior moment estimation in AIP with adaptive quadrature.

The marginal log-likelihood is obtained by combining adaptive quadrature integration over the item residuals $\epsilon_{\alpha i}^{(1)}$, $\epsilon_{\beta i}^{(1)}$, $\epsilon_{\alpha c}^{(2)}$, and $\epsilon_{\beta c}^{(2)}$ with Monte Carlo integration over the person trait value θ_p.

4. Empirical Illustration

It is clear from the work by Embretson (1999) and Daniel and Embretson (2010) that a combination of the item cloning and item attribute approach can be used for item generation. In this application, data from Daniel and Embretson (2010) will be analyzed with the AMIS model.

4.1. GRE Data

We refer to the data as GRE data because the items are similar to graduate record examination (GRE) test math items. All 405 participants were undergraduates at a large midwestern

university. Participants were randomly assigned to one of three parallel test forms. Twenty-two participants had incomplete test forms and were subsequently removed from the database. Thus, data from 383 participants (131 for Form 1, 128 for Form 2, and 123 for Form 3) were considered. The item structure used by Daniel and Embretson (2010) is one with two levels on the item side. Based on 36 items, two other sets of 36 items were generated, such that 36 item families were obtained, with three items each. All 36 item families are represented in each parallel form. The individual items define level 1 of the item structure, while the 36 item families define level 2.

Level-1 or Item Attributes Items within an item family are characterized by two item attributes: Equation Source and Number of Subgoals. Half of the families show a within difference with respect to the former and another half with respect to the latter.

Equation Source has three values: Given, where the required equation is presented in the item stem; Translate, where the required equation is given in words and must be translated into symbolic form; and Recall/Generate, where the required equation must be recalled from memory or developed uniquely for the problem. Examples of the three kinds of equation sources in a geometry problem to find an unknown side of a right triangle, are a diagram plus the Pythagorean theorem equation (Given); a diagram plus a statement of the Pythagorean theorem in words (Translate); and a diagram only (Recall/Generate). Two Helmert contrasts were specified for the attribute: Translate vs. Given and Recall/Generate vs. Other.

Number of Subgoals denotes the number of values that must be found prior to solving the main equation. Suppose that the central problem was to find the time needed for one vehicle to overtake another vehicle, given a specified difference in starting times. Number of Subgoals has three values: 0 subgoals, where the speed of each vehicle is specified; 1 subgoal, where the speed of one vehicle is given but the other must be computed from distance and time information; and 2+ subgoals, where the speed of each vehicle must be computed from distance and time information. Two Helmert contrasts were specified for Number of Subgoals: 0 vs. 1 Subgoal and 2+ vs. less Subgoals.

Level-2 or Item Category Attributes There are three category attributes, two of which are the attributes that are left unchanged in the offspring items. These are Equation Source when Number of Subgoals is varied in the offspring, and Number of Subgoals when Equation Source is varied in the offspring. The category attributes are the attributes whose values remain associated with the formulation and context of the original math problem. The third category attribute is the Relative Definition of Variable. Relative definition means that the unknown variables are defined relative to each other, thus requiring a solution of simultaneous equations. Relative Definition of Variables is coded as 1 vs. -1 for Relative vs. Not Relative.

Note that the category (level 2) attributes Number of Subgoals and Equation Source are the same variables as the item (level 1) attributes, and one of these variables takes the same value for the item category and each item within it. The reason for the distinction between the item attributes at level 1 and level 2 reflects the design of the items. The items were generated from item models. Item models in this study were items that had previously been found to have good psychometric properties. The level-2 attributes concern differences between item models. The specific items that were generated from the item models were designed to differ, reflecting differences in either Equation Source or Number of Subgoals. Thus, level-1 attributes concern differences between variants of item models. Whether the attributes with and without manipulation, corresponding to level 1 and level 2, respectively, have the same effect is an empirical issue. Note that it can also make sense to use similar explanatory variables for different levels in a multilevel regression model for persons within person groups.

The item structure as described implies that after the Helmert coding there are four two-valued item attributes (IA) and five two-valued category attributes (CA). The four item attributes

are: Given vs. Translate and Recall/Generate vs. Other (Source of Equation), and 0 vs. 1 Subgoal and 2+ Subgoals vs. Other (Number of Equations). The five category attributes are the same four variables, but now at level 2, plus Relative vs. Not Relative. Two item categories were excluded based on the item fit analysis by Daniel and Embretson (2010). All subsequent analyses were implemented with 34 item families. In addition to the 34 item families, 11 linking items were used. They belong to each of the three forms, such that the total number of items is 113 ($=34 \times 3 + 11$).

4.2. Method

The data were analyzed by Daniel and Embretson (2010) with fixed item attributes (FIA) and with fixed category (FC) effects, and thus without category attributes and without random item or category residuals. Because participants were randomly assigned to one of three equivalent test forms, the 11 linking items are used to link the scales of the three forms. These 11 items are always modeled with fixed effects, constrained equal across test forms.

We will perform an analysis with four models:

1. *The FI Model.* The model has only one level and individual fixed item effects: their difficulty and discrimination. The FI model does not explain the item parameters and does not have predictive value for items generated in the future.
2. *The FCA&FIA&RI Model.* This model also has only one level. All attributes, item attributes, and category attributes are used, but only one random residual, at the item level, is added for item discriminations and difficulties. Compared with the full AMIS model, the FCA&FIA&RI model relies on the same attributes as explanatory variables, but the multilevel structure is ignored. A simulation study will investigate the effects of ignoring the second level, even when keeping its explanatory variables. The FCA&FIA&RI model as applied is

$$\eta_{pic} = \left[(\mu_a + a_i)\cdot L_i + \left(\mu_\alpha + \sum_t \delta_{\alpha t} R_{ct} + \sum_d \gamma_{\alpha d} Q_{id} + \epsilon^{(1)}_{\alpha i}\right)(1 - L_i)\right]\cdot \theta_p - \left[(\mu_b + b_i)\cdot L_i + \left(\mu_\beta + \sum_t \delta_{\beta t} R_{ct} + \sum_d \gamma_{\beta d} Q_{id} + \epsilon^{(1)}_{\beta i}\right)(1 - L_i)\right], \quad (9)$$

   where L_i is an indicator for linking items, taking the value 1 if item i is a linking item and 0 otherwise, μ_a and μ_b are means of linking items, and μ_α and μ_β are intercepts of nonlinking items. Constraints, $\sum a_i = 0$ and $\sum b_i = 0$, are used for the fixed item discriminations and difficulties of the linking items, respectively. Helmert contrasts are used for R_{ct} and Q_{id}.
3. *The FCA&RC&FIA&RI Model or Full AMIS Model.* The model has two levels, with fixed item attribute effects at level 1 and fixed category attribute effects at level 2, with random item residuals for the unexplained item variation, and random category residuals for the unexplained category variation. This multilevel regression model for the item structure is analogous to a multilevel regression model for respondents nested in groups such as schools. The FCA&RC&FIA&RI model as applied is

$$\eta_{pic} = \left[(\mu_a + a_i)\cdot L_i + \left(\mu_\alpha + \sum_t \delta_{\alpha t} R_{ct} + \sum_d \gamma_{\alpha d} Q_{id} + \epsilon^{(1)}_{\alpha i} + \epsilon^{(2)}_{\alpha c}\right)(1 - L_i)\right]\cdot \theta_p - \left[(\mu_b + b_i)\cdot L_i + \left(\mu_\beta + \sum_t \delta_{\beta t} R_{ct} + \sum_d \gamma_{\beta d} Q_{id} + \epsilon^{(1)}_{\beta i} + \epsilon^{(2)}_{\beta c}\right)(1 - L_i)\right]. \quad (10)$$

4. *The RC&RI Model.* The model has random effects at two levels but no attribute effects. The model is considered in order to have an idea of the variation of each level when the attributes are not invoked to explain some of the variance. This is analogous to the "unconditional model" in a multilevel model for respondents nested in groups. Although the residual variances from the full AMIS model (FCA&RC&FIA&RI) can strictly speaking not be compared with the variances in the RC&RI model, a comparison of these variance estimates can nevertheless give an approximative indication of the proportion of variance explained. The RC&RI model as applied is

$$\begin{aligned}\eta_{pic} = &\left[(\mu_a + a_i)\cdot L_i + \left(\mu_\alpha + \varepsilon_{\alpha i}^{(1)} + \varepsilon_{\alpha c}^{(2)}\right)(1 - L_i)\right]\cdot \theta_p \\ &- \left[(\mu_b + b_i)\cdot L_i + \left(\mu_\beta + \varepsilon_{\beta i}^{(1)} + \varepsilon_{\beta c}^{(2)}\right)(1 - L_i)\right]. \end{aligned} \tag{11}$$

The FCA&FIA&RI and RC&RI models are nested in the full AMIS model, but the FI model is not. Parameters of the FI model were estimated using marginal maximum likelihood estimation implemented in **gllamm** and those of the FIA&RI, FCA&RC&FIA&RI, and RC&RI models were estimated using the AIP algorithm with two or three wings. Fifteen-point adaptive quadrature was used throughout. For the evaluation of the log-likelihood, the person abilities were sampled and the 4-dimensional integrals over the item and category residuals were evaluated using adaptive quadrature with 8 quadrature points per dimension for each random draw of the person abilities. The average over the draws of the ability estimates was then formed, thus using Monte Carlo integration over persons and adaptive quadrature over item and category residuals.

4.3. Results

The goodness of fit of the four models was compared based on Akaike's (1974) information criterion (AIC) and Schwarz's (1978) Bayesian information criterion (BIC) as summarized in Table 1. The appropriate use of information criteria in latent variable models is an ongoing area of research (e.g., Vaida & Blanchard, 2005), and we use AIC and BIC indices informally in this study. For the BIC the sample size was taken to be the number of respondents. The full AMIS model (FCA&RC&FIA&RI) fits better than the two-level model without attributes (RC&RI), according to the AIC (19670 vs. 19697) but not according to the BIC (19852 vs. 19807). The full AMIS model also fits better than the one-level model with all attributes and a random item residual (FCA&FIA&RI), with an AIC of 19670 vs. 19844, and a BIC of 19852 vs. 20017. Finally, when compared with the model with only fixed item effects (FI), the commonly used model in practice, the full AMIS model fits better according to the BIC (19852 vs. 19993) but not according to the AIC (19670 vs. 19101). Because of its comparatively good fit, its explanatory nature and two-level item structure, the two-level model with attributes and random item effects (FCA&RC&FIA&RI) is chosen as the estimated model to discuss further. Without a substantial penalty for the number of parameters, the goodness of fit of a fully fixed item model is per definition better, just as the fully fixed person effect (as estimated by joint maximum likelihood estimation) always has a better log-likelihood than the model estimated with random person

TABLE 1.
Model fit comparisons for GRE data.

Models	Levels	Effects	Attributes	Npar	Log-likelihood	AIC	BIC
FI	One-Level	Fixed	No	226	−9324.5	19101	19993
FCA&FIA&RI	One-Level	Random	Yes	44	−9877.8	19844	20017
FCA&RC&FIA&RI	Two-Level	Random	Yes	46	−9789.2	19670	19852
RC&RI	Two-Level	Random	No	28	−9820.3	19697	19807

TABLE 2.
FCA&FIA&RI and FCA&RC&FIA&RI model estimates for GRE data.

Item predictors	FCA&FIA&RI		FCA&RC&FIA&RI	
	Est	SE	Est	SE
Fixed effects				
Intercept				
μ_α	0.800*	0.049	0.875*	0.051
μ_β	0.623*	0.060	0.672*	0.059
Item-Level				
for item discriminations				
$\gamma_{\alpha 1}$[Given vs. Translate]	0.007	0.038	0.010	0.044
$\gamma_{\alpha 2}$[Recall/Generate vs. Other]	−0.051	0.032	−0.010	0.022
$\gamma_{\alpha 3}$[0 vs. 1 Subgoal]	−0.104	0.061	−0.069	0.068
$\gamma_{\alpha 4}$[2+ Subgoals vs. Other]	−0.059*	0.030	−0.077*	0.035
for item difficulties				
$\gamma_{\beta 1}$[Given vs. Translate]	−0.085	0.089	−0.067	0.098
$\gamma_{\beta 2}$[Recall/Generate vs. Other]	0.189	0.120	0.007	0.056
$\gamma_{\beta 3}$[0 vs. 1 Subgoal]	0.433*	0.205	0.296	0.179
$\gamma_{\beta 4}$[2+ Subgoals vs. Other]	0.295*	0.104	0.305*	0.111
Category-Level				
for item discriminations				
$\delta_{\alpha 1}$[Given vs. Translate]	0.336*	0.122	0.311*	0.128
$\delta_{\alpha 2}$[Recall/Generate vs. Other]	−0.110	0.065	−0.150*	0.062
$\delta_{\alpha 3}$[0 vs. 1 Subgoal]	−0.001	0.094	−0.059	0.095
$\delta_{\alpha 4}$[2+ Subgoals vs. Other]	0.004	0.045	−0.047	0.046
$\delta_{\alpha 5}$[Relative vs. Absolute]	0.126	0.119	0.124	0.129
for item difficulties				
$\delta_{\beta 1}$[Given vs. Translate]	−0.373	0.228	−0.194	0.239
$\delta_{\beta 2}$[Recall/Generate vs. Other]	0.226	0.135	0.356*	0.143
$\delta_{\beta 3}$[0 vs. 1 Subgoal]	−0.023	0.119	0.094	0.240
$\delta_{\beta 4}$[2+ Subgoals vs. Other]	−0.027	0.101	0.124	0.104
$\delta_{\beta 5}$[Relative vs. Absolute]	0.133	0.280	0.042	0.276
Random effects				
Item-Level				
$\psi_\alpha^{(1)}$[Variance]	0.150		0.057	
$\psi_\beta^{(1)}$[Variance]	1.270		1.103	
Category-Level				
$\psi_\alpha^{(2)}$[Variance]	–		0.051	
$\psi_\beta^{(2)}$[Variance]	–		0.082	

*: Significance based on p-value $< .05$ for fixed effects
–: Not modelled

effects (as estimated by marginal maximum likelihood estimation). The issue is rather the possibly poor estimation of the item parameters when the multilevel item structure is not taken into account.

Table 2 shows the parameter estimates (Est) and their standard errors (SE) for the FCA&FIA&RI and the full AMIS model. Wald-based inferences tend to perform poorly for

variance parameters and therefore SEs are not reported for the variance estimates. The estimates of the item attribute effects are very similar, but the level-1 variance estimates are smaller in the full AMIS model, accompanied by generally somewhat larger estimated standard errors for the category attribute effects, as would be expected. This is a first indication of what it means to ignore the second level.

In the FCA&RC&FIA&RI model, the estimated effects of the item attributes on item discrimination are not too strong. The only significant coefficient based on a Wald test at the 5 % level is for 2+ Subgoals vs. Other, estimated as -0.077, implying that the effect on the discrimination is negative. The effect is rather small given that the estimate of the intercept is 0.875. The unexplained variance is estimated as 0.057. The estimated effects of the item attributes on the item difficulty are also not strong. The only significant effect stems from the same attribute, with an estimated coefficient of 0.305. Apparently, two or more subgoals make the items more difficult while it makes them also somewhat less discriminative. The difficulty intercept is estimated as 0.672 and the unexplained level-1 variance as 1.103, which is rather large.

The estimated effects of the category attributes are again not strong. The following category attributes have significant effects on the discrimination: The coefficients of Given vs. Translate and Recall/Generate vs. Other, are estimated as 0.311 and -0.150, respectively. The unexplained level-2 discrimination variance is estimated as 0.051. There is only one significant effect of the category attributes on difficulty: Recall/Generate vs. Other. Recall/Generate seems to make item families more difficult, with an estimated coefficient of 0.356. Apparently, item families that require the recall or generation of an equation are both more difficult and less discriminating. Finally, Relative Definition of Variables does not seem to have a significant effect. The unexplained level-2 difficulty variance is estimated as 0.082.

The explanatory power of the attributes can be approximately investigated through a comparison with the fixed item effects (FI) model. When the item discriminations and difficulties are predicted on the basis of the level-1 and level-2 attribute effects, the correlations with the estimated FI discriminations and difficulties are 0.438 and 0.456, respectively. When only the level-1 attributes are taken into account, the correlations with the FI discriminations and difficulties are 0.223 and 0.274. When only the level-2 attributes are taken into account, the correlations with the FI discriminations and difficulties are 0.386 and 0.366.

The explanatory value of the attributes can also be assessed by a comparison between the full AMIS model (FCA&RC&FIA&RI) and the corresponding model without attribute effects (RC&RI) as suggested by Snijders and Bosker (1994) for hierarchical linear models. The percentage of the variance explained by the attributes is 85 % for discrimination (total estimated residual variance of 0.108 in AMIS and 0.724 in RC&RI) and 15 % for difficulty (total estimated residual variance of 1.185 in AMIS and 1.391 in RC&RI). These results indicate that the attributes have a good explanatory (and thus predictive) value for the discriminations but much less for the difficulties.

5. Simulation Study

The consequences of ignoring the dependency stemming from the item categories are investigated through a simulation study comparing the results of FCA&FIA&RI and FI with those of FCA&RC&FIA&RI. Previous studies reported that less accurate item parameter estimates (or predictions) are obtained when the dependency within item categories is ignored (Bradlow et al., 1999; Glas & van der Linden, 2003; Sinharay et al., 2003). However, these studies did not include the full AMIS model.

The estimates of the full AMIS model as reported in Table 2 were considered as true parameters in order to generate 10 data sets. The full AMIS model and its single-level counterpart, the

TABLE 3.
Simulation study results based on FCA&FIA&RI and FCA&RC& FIA&RI model estimates from GRE data.

Item predictors	FCA&FIA&RI		FCA&RC&FIA&RI	
	Bias	RMSE	Bias	RMSE
Fixed effects				
Intercept				
μ_α	−0.060*	0.071	−0.010	0.060
μ_β	−0.053	0.081	0.001	0.002
Item-Level				
for item discriminations				
$\gamma_{\alpha 1}$[Given vs. Translate]	−0.090*	0.108	−0.004	0.088
$\gamma_{\alpha 2}$[Recall/Generate vs. Other]	−0.042	0.017	−0.001	0.009
$\gamma_{\alpha 3}$[0 vs. 1 Subgoal]	0.056*	0.090	0.024*	0.085
$\gamma_{\alpha 4}$[2+ Subgoals vs. Other]	−0.099*	0.111	0.036*	0.060
for item difficulties				
$\gamma_{\beta 1}$[Given vs. Translate]	0.025	0.037	0.008	0.018
$\gamma_{\beta 2}$[Recall/Generate vs. Other]	0.089*	0.091	−0.019*	0.042
$\gamma_{\beta 3}$[0 vs. 1 Subgoal]	0.106*	0.194	0.018	0.049
$\gamma_{\beta 4}$[2+ Subgoals vs. Other]	−0.099*	0.143	0.001	0.020
Category-Level				
for item discriminations				
$\delta_{\alpha 1}$[Given vs. Translate]	−0.032*	0.039	−0.013	0.016
$\delta_{\alpha 2}$[Recall/Generate vs. Other]	−0.021*	0.020	0.002	0.008
$\delta_{\alpha 3}$[0 vs. 1 Subgoal]	−0.025*	0.035	0.018*	0.028
$\delta_{\alpha 4}$[2+ Subgoals vs. Other]	0.019	0.027	0.012	0.025
$\delta_{\alpha 5}$[Relative vs. Absolute]	0.018	0.021	0.016	0.020
for item difficulties				
$\delta_{\beta 1}$[Given vs. Translate]	−0.018	0.029	0.007	0.019
$\delta_{\beta 2}$[Recall/Generate vs. Other]	−0.092*	0.086	−0.032*	0.051
$\delta_{\beta 3}$[0 vs. 1 Subgoal]	−0.033	0.065	−0.023	0.036
$\delta_{\beta 4}$[2+ Subgoals vs. Other]	−0.039*	0.034	−0.011	0.024
$\delta_{\beta 5}$[Relative vs. Absolute]	0.090*	0.021	0.007	0.010
Random effects				
Item-Level				
$\psi_\alpha^{(1)}$[Variance]	0.091*	0.241	−0.010	0.017
$\psi_\beta^{(1)}$[Variance]	0.138*	0.154	−0.067	0.084
Category-Level				
$\psi_\alpha^{(2)}$[Variance]	–	–	−0.009	0.018
$\psi_\beta^{(2)}$[Variance]	–	–	−0.018	0.023

*: Significantly different from 0 at the 5 % level
–: Not modelled

FCA&FIA&RI model, were fit to the 10 simulated data sets using AIP with fifteen-point adaptive quadrature. Monte Carlo errors of the bias and RMSE, estimated as suggested by Koehler, Brown, and Haneuse (2009) were less than 0.011 for the bias and less than 0.013 for the RMSE. This is a reassuring finding given the rather small number of replications. Table 3 reports the estimated bias and root mean square error (RMSE) of the regression coefficients and variances of

the random item and category residuals for discrimination and difficulty. The bias and RMSE are clearly larger for the FCA&FIA&RI model than for the full AMIS model. The variances of the random item effects are overestimated in the FCA&FIA&RI model as a consequence of ignoring the multilevel structure.

Given the parameter estimates, the correlations between the true and predicted item parameters were obtained, where the predicted values are $(\widehat{\mu}_\alpha + \sum_d \widehat{\gamma}_{\alpha d} \cdot Q_{id} + \tilde{\varepsilon}^{(1)}_{\alpha i} + \tilde{\varepsilon}^{(2)}_{\alpha c})$ for the discriminations and $(\widehat{\mu}_\beta + \sum_d \widehat{\gamma}_{\beta d} \cdot Q_{id} + \tilde{\varepsilon}^{(1)}_{\beta i} + \tilde{\varepsilon}^{(2)}_{\beta c})$ for the difficulties. The average correlations between true parameters and predictions across the ten replications for the full AMIS and FCA&FIA&RI models are 0.955 and 0.884, respectively, for discriminations, and 0.997 and 0.902, respectively, for difficulties. These results indicate that ignoring the multilevel item structure leads to suboptimal prediction.

The predictions of the discriminations and difficulties from the AMIS model and the corresponding estimates from the FI model were also evaluated by obtaining the average estimated bias and the square root of the average MSE across items. To compare the bias for item parameters between the AMIS and FI models, the average difference between each difficulty (or discrimination) estimate and the corresponding generating value was averaged over items for each replicate dataset. Paired t tests, with 9 degrees of freedom, found significant differences for both the discriminations and the difficulties at the 5 % level. For the AMIS model, the average bias estimates were 0.032 and -0.057 for discriminations and difficulties, respectively, compared with 0.131 and -0.155 for the FI model. The biases were statistically significant using one-sample t tests at the 5 % level for all items in the FI model but not for all items (10 % of all items) in the AMIS model. The square root of the average MSEs for discriminations and difficulties in the full AMIS model were 0.257 and 0.201, respectively, compared with 0.392 and 0.384 for the FI model. These results show that ignoring the multilevel item structure and using fixed item effects lead to less accurate item estimates.

6. Discussion

Based on the variety of approaches and models available, the AMIS model was proposed, starting from earlier work by Geerlings et al. (2011), Glas and van der Linden (2003), and Sinharay et al. (2003) on the one hand, and Embretson (1999) and Daniel and Embretson (2010) on the other hand. The model is a way to integrate different approaches and to indicate modeling possibilities that have not yet been tried, such as models with random category effects in addition to random item effects. For example, when category attributes are used as in the model of Geerlings et al. (2011), it can pay off to include a random residual at the level of the item categories.

A basic feature of the AMIS models is that random item residuals are used. Item response theory (IRT) parameter estimation with random item parameters is not common in the domain of IRT (Baker & Kim, 2004) and MCMC estimation is almost always used when items are treated as random (Fox, 2010). Random item effects do not pose a problem as such for a fully Bayesian approach, but as surveyed in Cho and Suh (2012), there is no consensus which distribution should be specified for the discrimination parameters and their variance. AIP with adaptive quadrature does not require priors for model parameters. However, in line with maximum likelihood estimation methods, a downward bias can be found with AIP if the number of clusters is small (Cho & Rabe-Hesketh, 2011). In linear models, this problem is addressed by using restricted maximum likelihood (REML) estimation (Patterson & Thompson, 1971). Unfortunately, the REML concept cannot be directly applied to generalized linear mixed models (and item response model), although there are some approaches described in the literature (e.g., Bellio & Brazzale, 2011; Breslow & Clayton, 1993; McGilchrist, 1994; Noh & Lee, 2007).

The simulation study was based on the empirical study. Results show that the parameters of the full AMIS model (FCA&RC&FIA&RI) are recovered well using AIP with adaptive quadrature. In addition, the study shows that ignoring the multilevel item structure has undesired effects for the quality of the estimates (bias, RMSE). Further simulation studies are required to investigate the systematic performance of the estimation methods with respect to the number of items, the number of categories, the different structure of attributes, and the degree of dependency of items within the same item category.

In the following, some remaining modeling and testing issues are discussed. When the item difficulties and the item discriminations are modeled with a random component, one may consider the possibility that they are correlated. However, in a model with sufficient explanatory power, one may expect the residuals to be independent.

In the application, the FI model was used as an alternative model. The number of examinees in the application is relatively small to estimate all item parameters of the FI model under marginal maximum likelihood estimation. Even though there were no convergence problems and no extremely large discrimination estimates in the application, it is possible that this kind of problem interferes with the quality of the results in our application.

It may be useful to test the variances in the AMIS model. One major concern when testing variances is that the null hypothesis is situated on the boundary of the parameter space, so that the asymptotic distribution of the standard likelihood ratio test statistic is no longer Chi-quare (e.g., Stram and Lee, 1994, 1995). The asymptotic distribution for testing variances for one or more random effects can be found in Gurieroux, Holly, and Monfort (1982) and Verbeke and Molenberghs (2003).

The AMIS model described in this paper can be extended in two ways. First, the AMIS model can be extended to polytomous responses. The item difficulty can be decomposed into item locations and category thresholds in polytomous item response models. For polytomous AMIS models, a random residual can be added to either the item locations or category thresholds or both. Second, AMIS models can be extended to be multidimensional. For example, the item attributes relevant to items for a given dimension can be used to explain or predict variations in item effects, depending on the dimension in question.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Albers, W., Does, R.J.M.M., Imbos, T., & Janssen, M.P.E. (1989). A stochastic growth model applied to repeated test of academic knowledge. *Psychometrika*, *54*, 451–466.

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

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale: Erlbaum.

Bejar, I.I. (2012). Item generation: implications for a validity argument. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation*, New York: Taylor & Francis.

Bejar, I.I., Lawless, R.R., Morley, M.E., Wagner, M.E., Bennett, R.E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, *2*, 3–28.

Bellio, R., & Brazzale, A.R. (2011). Restricted likelihood inference for generalized linear models. *Statistics and Computing*, *21*, 173–183.

Birnbaum, A. (1968). Test scores, sufficient statistics, and the information structures of tests. In L. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 425–435). Reading: Addison-Wesley.

Bock, R.D., & Schilling, S.G. (1997). High-dimensional full-information item factor analysis. In M. Berkane (Ed.), *Latent variable modelling and applications to causality* (pp. 164–176). New York: Springer.

Bormuth, J.R. (1970). *On the theory of achievement test items*. Chicago: University of Chicago Press.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

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

Breslow, N.E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, *82*, 81–91.

Breslow, N.E. (2004). Whither PQL? In D.Y. Lin & P.J. Heagerty (Eds.), *Proceedings of the second seattle symposium in biostatistics: analysis of correlated data* (pp. 1–22). New York: Springer.

Browne, W.J., & Draper, D. (2006). A comparison of Bayesian and likelihood methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473–514.

Chaimongkol, S., Huffer, F.W., & Kamata, A. (2006). A Bayesian approach for fitting a random effect differential item functioning across group units. *Thailand Statistician*, *4*, 27–41.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, *55*, 12–25.

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

Cho, S.-J., Athay, M., & Preacher, K.J. (2013). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical & Statistical Psychology*, *66*, 353–381.

Cho, S.-J., Gilbert, J.K., & Goodwin, A.P. (2013). Explanatory multidimensional multilevel random item response model: an application to simultaneous investigation of word and person contributions to multidimensional lexical quality. *Psychometrika*, *78*, 830–855.

Clayton, D.G., & Rasbash, J. (1999). Estimation in large crossed random-effect models by data augmentation. *Journal of the Royal Statistical Society Series A*, *162*, 425–436.

Daniel, R.C., & Embretson, S.E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, *34*, 348–364.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Jong, M.G., Steenkamp, J.B.E.M., & Fox, J.-P. (2007). Relaxing cross-national measurement invariance using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278.

De Jong, M.G., Steenkamp, J.B.E.M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: a global investigation. *Journal of Marketing Research*, *45*, 104–115.

De Jong, M.G., & Steenkamp, J.B.E.M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large-scale cross-cultural research. *Psychometrika*, *75*, 3–32.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychological Methods*, *3*, 300–396.

Embretson, S.E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, *64*, 407–433.

Embretson, S.E. (2010). Cognitive design systems: a structural modelling approach applied to developing a spatial abtiliy test. In S.E. Embretson (Ed.), *Measuring psychological constructs: advances in model-based approaches* (pp. 247–273). Washington: American Psychological Association.

Embretson, S.E., & Daniel, R.C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*, *50*, 328–344.

Embretson, S.E., & Gorin, J.S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.

Embretson, S.E., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics: psychometrics* (Vol. 26, pp. 747–768). North Holland: Elsevier.

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

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

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: a random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*, 432–457.

Freund, Ph.A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, *32*, 195–210.

Geerlings, H., Glas, C.A.W., & van der Linden, W.J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*, 337–359.

Gierl, M., & Haladyna, T. (2012). *Automatic item generation*. New York: Taylor & Francis.

Gierl, M., & Lai, H. (2012). Using weak and strong theory to create item models for automatic item generation: some practical guidelines with examples. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation*, New York: Taylor & Francis.

Gierl, M.J., Zhou, J., & Alves, C.B. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning, and Assessment*, *7*, 1–51.

Glas, C.A.W., & van der Linden, W.J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247–261.

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. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, *78*, 45–51.

Gorin, J. (2005). Manipulating processing difficulty of reading comprehension questions: the feasibility of verbal item generation. *Journal of Educational Measurement*, *42*, 351–373.

Gurieroux, C., Holly, A., & Monfort, A. (1982). Likelihood ratio test, Wald test, and Kuhn–Tucker test in linear models with inequality constraints on the regression parameters on the regression parameters. *Econometrica*, *50*, 63–80.

Holling, H., Bertling, J.P., & Zeuch, N. (2009). Probability word problems: automatic item generation and LLTM modelling. *Studies in Educational Evaluation*, *35*, 71–76.

Irvine, S.H. & Kyllonen, P. (Eds.) (2002). *Item generation for test development*. Mahwah: Erlbaum.

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.

Janssen, R., Schepers, J., & Perez, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: a generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, *52*, 5066–5074.

Johnson, M.S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement*, *29*, 369–400.

Karim, M.R., & Zeger, S.L. (1992). Generalized linear models with random effects: Salamander mating revisited. *Biometrics*, *48*, 631–644.

Klein Entink, R.H., Fox, J.-P., & van der Linden, W.J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Klein Entink, R.H., Kuhn, J.-T., Hornke, L.F., & Fox, J.-P. (2009b). Evaluating cognitive theory: a joint modeling approach using responses and response times. *Psychological Methods*, *14*, 54–75.

Koehler, E., Brown, E., & Haneuse, S. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *American Statistician*, *63*, 155–162.

Lee, Y., & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society Series B*, *58*, 619–678.

Lee, Y., & Nelder, J.A. (2006). Double-hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society Series C*, *55*, 1–29.

Lin, X., & Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, *91*, 1007–1016.

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society Series B*, *56*, 61–69.

Millman, J., & Westman, R.S. (1989). Computer assisted writing of achievement test items: toward a future technology. *Journal of Educational Measurement*, *26*, 177–190.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195.

Mislevy, R.J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, *12*, 281–296.

Natarajan, R., & Kass, R.E. (2000). Reference Bayesian methods for generalized linear mixed model. *Journal of the American Statistical Association*, *95*, 227–237.

Noh, M., & Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, *98*, 896–915.

Patterson, H.D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*, 545–554.

Pinheiro, J.C., & Bates, D.M. (1995). Approximation to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational Graphics and Statistics*, *4*, 12–35.

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

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station: Stata Press.

Rasbash, J., & Browne, W.J. (2007). Non-hierarchical multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 333–336). New York: Springer.

Raudenbush, S.W., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*, 141–157.

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*, 73–89.

Rodriguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case study. *Journal of the Royal Statistical Society Series A*, *164*, 339–355.

Roid, G.H., & Haladyna, T.M. (1982). *Toward a technology of test-item writing*. New York: Academic.

Schilling, S., & Bock, R.D. (2005). High dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Scrams, D.J., Mislevy, R.J., & Sheehan, K.M. (2002). *An analysis of similarities in item functioning within antonym and analogy variant families (RR-02-13)*. Princeton: Educational Testing Service.

Sinharay, S., Johnson, M.S., & Williamson, D.M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics*, *28*, 295–313.

Snijders, T.A.B., & Bosker, R.J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*, 342–363.

Soares, T.M., Gonçalvez, F.B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*, 348–377.

Stram, D.O., & Lee, J.W. (1994). Variance components testing in the longitudinal mixed effect model. *Biometrics*, *50*, 1171–1177.

Stram, D.O., & Lee, J.W. (1995). Correction to: variance components testing in the longitudinal mixed-effects model. *Biometrics*, *51*, 1196.

Tanner, M.A., & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.
Tierney, L., & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.
Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *Biometrika*, *92*, 351–370.
van der Linden, W.J., Klein Entink, R.H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347.
Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, *59*, 254–262.
Wainer, H., Bradlow, E.T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Manuscript Received: 10 JUN 2011
Published Online Date: 12 DEC 2013