PSYCHOMETRIKA—VOL. 76, NO. 2, 337–359
APRIL 2011
DOI: 10.1007/S11336-011-9204-X

MODELING RULE-BASED ITEM GENERATION

HANNEKE GEERLINGS AND CEES A.W. GLAS
UNIVERSITY OF TWENTE

WIM J. VAN DER LINDEN
CTB/MCGRAW-HILL

An application of a hierarchical IRT model for items in families generated through the application of different combinations of design rules is discussed. Within the families, the items are assumed to differ only in surface features. The parameters of the model are estimated in a Bayesian framework, using a data-augmented Gibbs sampler. An obvious application of the model is computerized algorithmic item generation. Such algorithms have the potential to increase the cost-effectiveness of item generation as well as the flexibility of item administration. The model is applied to data from a non-verbal intelligence test created using design rules. In addition, results from a simulation study conducted to evaluate parameter recovery are presented.

Key words: hierarchical modeling, item generation, item response theory, Markov chain Monte Carlo method.

1. Introduction

One of the main reasons why automated item generation has gained interest lately is the need for large item pools to further flexibility in test administration while avoiding overexposure of the items. When done manually, item writing can be a costly and time-consuming endeavor. However, given a well-specified set of rules, a computer can generate a large pool of items in a negligible amount of time. An additional advantage of automated item generation is the availability of precise information about how the items have been constructed. For instance, the information can be used as a check on the validity of the test. In the present paper, a hierarchical item response theory (IRT) model incorporating information about the item-design rules is used to analyze a dataset consisting of responses to rule-based generated items. The parameters of the model are estimated in a Bayesian fashion, using a data-augmented Gibbs sampler.

The model is developed for a combination of two methods of automated item generation. The first method is generation based on cognitive analysis of the item domain. The results from the analysis are then used to devise rules for the generation of new items (Embretson, 1999). Irvine (2002) introduced the term "radicals" to refer to such rules. An example of a radical is whether or not Bayes' rule has to be applied to solve a statistics item. Radicals can be used to automate item generation. In addition, the radicals can be assumed to be important determinants of item difficulty. A psychometric model accounting for the effects of radicals is the linear logistic test model (LLTM; Fischer, 1973; Freund, Hofer, & Holling, 2008; Holling, Bertling, & Zeuch, 2009). This model decomposes the difficulty parameter of the Rasch (1960) model into a linear combination of the effects of radicals. An error term can be added to the model to make it less restrictive.

The second method is item cloning. The goal of item cloning is to generate a set or family of items that look different but are generated by the same combination of radicals. The families

Requests for reprints should be sent to Hanneke Geerlings, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: h.geerlings@gw.utwente.nl

are created from parent items for the combinations of radicals by changing some of their surface features. Irvine (2002) refers to these features as "incidentals". Incidentals are not assumed to influence the difficulty of the items in any systematic way; their only goal is to ensure that items within a family are sufficiently different to avoid solving them just by remembering earlier solutions. For the earlier example of a statistics item, an incidental could be a context story with information irrelevant to the formal statistical problem. Incidentals can be produced, for example, with the help of replacement sets for some of the insignificant elements of the parent items (Hively, Patterson, & Page, 1968; Millman & Westman, 1989; Osburn, 1968; Roid & Haladyna, 1982), by transforming their text by means of linguistic rules (Bormuth, 1970), or by applying other natural language generation techniques. A psychometric model for this approach is the hierarchical model proposed by Glas and van der Linden (2001, 2003; see also Sinharay, Johnson, & Williamson, 2003; Glas, van der Linden, & Geerlings, 2010). This model, which will be referred to as the item cloning model (ICM), assumes that the parameters of the individual items are a combination of family parameters with a random component to allow for the unsystematic variation caused by incidentals. In principle, if the family parameters have been estimated from a previous sample of items with enough precision, newly generated items would not have to be calibrated at all, because their parameters can simply be assumed to be drawn from the known family distributions.

Ideally, a system for automated item generation based on these two methods can produce a large collection of item families. Within each family, similarity among items is caused by the use of the same radicals whereas dissimilarities would be the result of incidentals only. In the present paper, we combine the ICM with an LLTM-like structure for the expected value of the item difficulty parameters for each family. The structure decomposes the mean family difficulty into separate effects for each of its radicals.

The model, which will be labeled the *linear item cloning model* (LICM), is discussed in more detail in the next section. In the third section, an empirical study using a dataset from a non-verbal intelligence test is presented to show how the model can be applied in practice. Furthermore, a simulation study was conducted to investigate the effect of different factors in the sampling design on the recovery of the model parameters during calibration. The results from this study will be discussed in the fourth section. The article concludes with a discussion of practical applications and future research on the model.

2. Modeling Approach

In IRT-based item calibration, person parameters are often considered random to represent the fact that the items are calibrated using data from a random sample of persons. In the current setting, items are treated as random as well, because they can be considered random instantiations ("clones") from their respective families. Therefore, to calibrate item families, it seems natural to model both the person and item parameters as random. The resulting model is a crossed-random effects model. Crossed-random effects models are difficult, but not impossible, to estimate in a frequentistic framework (see van den Noortgate, De Boeck, & Meulders, 2003; Glas & van der Linden, 2003; Cho & Rabe-Hesketh, 2011). However, component-wise estimation in the form of Gibbs sampling from the posterior distributions of the parameters reduces the estimation into manageable pieces. Gibbs sampling of the parameters of the ICM has been considered in Glas and van der Linden (2001) and Sinharay et al. (2003). A Gibbs sampler for a similar model, with the correlations between the item parameters restricted to zero, was proposed by Janssen, Tuerlinckx, Meulders, and De Boeck (2000) in the context of criterion-referenced measurement.

In the present paper, we do not only wish to account for item-cloning effects in the model, but also for item-generation rules that are hypothesized to have a fixed effect on the item difficulties.

The LLTM (Fischer, 1973) was one of the first examples of adding explanatory variables to otherwise descriptive models. The model was later extended to account for residual variance (random-effects LLTM; Janssen, Schepers, & Peres, 2004) and for random weights (random-weights LLTM; Rijmen & De Boeck, 2002). Explanatory models have been described in a more general nonlinear mixed modeling framework by Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003), and De Boeck and Wilson (2004).

To summarize, the LICM is a hierarchical IRT model with higher-level explanatory variables accounting for the effects of both item cloning and item generation rules. Fox and Glas (2001) presented a Gibbs sampler for a multilevel model with explanatory variables for the person parameters. In the present paper, the focus will be on explanatory variables for the item families. Extending the ICM by Glas and van der Linden (2003) with a linear structure on the family difficulty parameters has several advantages. First, the model can be used to check the theory used to generate the items, and thereby function as a quality control mechanism. In this regard, the model fits into the frameworks of assessment engineering (AE; Luecht, 2009) and evidence-centered design (ECD; Mislevy & Levy, 2007). Both frameworks share the point of view of assessment as a process of obtaining evidence about the ability of a test taker. Items developed according to a cognitive model, that is, a model with the cognitive steps a test taker has to take to solve the item, can provide such evidence. In doing so, psychometric models are used in a confirmatory manner—to test hypotheses provided by the cognitive model. For instance, methods for model comparison and model fit could be used to investigate whether the radicals properly explain the difficulty of a family, and whether the incidentals only have a random effect on the difficulty of the items (see Section 3.2). Second, using the information provided by the item-design process, the estimation of the difficulty of a particular family can borrow strength from data available for the other families (see Section 4.2). Finally, the model supports item generation on-the-fly; that is, test administration in which the items are sampled from calibrated families in real time and ability is estimated using the family parameters.

2.1. Response Model

Consider $f = 1, \ldots, F$ item families, with family f consisting of item $i_f = 1, \ldots, I_f$. In total, there are K items. The families are identified by combinations of radicals $r = 1, \ldots, R$. Each of $n = 1, \ldots, N$ persons is administered a subset of the K items, resulting in a response vector with realizations of response variables $U_{i_f n} = \{0, 1\}$ for every person n. Missing responses created by this design are considered as missing at random. Therefore, for convenience, and without loss of generality, we will not make the design explicit in the notation. Furthermore, it will be assumed that each item sampled from a family is administered to more than one test taker. Figure 1 offers a simplified representation of the design of an item pool with algorithmically generated items.

2.1.1. First-Level Model The first-level model specifies the probability of a person giving a correct response on an item as

$$p(U_{i_f n} = 1 | \theta_n, a_{i_f}, b_{i_f}, \gamma_{i_f}) = \gamma_{i_f} + (1 - \gamma_{i_f}) \Phi\big[a_{i_f}(\theta_n - b_{i_f})\big]. \tag{1}$$

This is the three-parameter normal-ogive (3PNO) model in which a_{i_f}, b_{i_f}, and γ_{i_f} are the item discrimination, difficulty and guessing parameters, respectively, θ_n is the person parameter, and $\Phi(\cdot)$ is the cumulative normal density function.

Alternatively, the 2PNO (the 3PNO without the guessing parameter) can be used as the first-level model. Note that Glas and van der Linden (2003) originally presented the ICM with the three-parameter logistic (3PL) model as the first-level model. However, the normal-ogive link function has the advantage of easy sampling from the conditional posterior distributions.

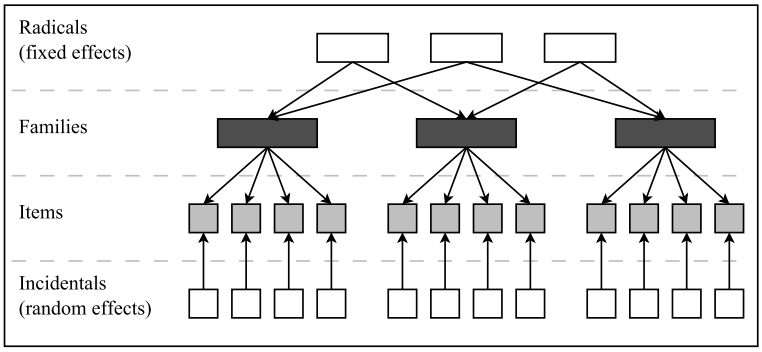

FIGURE 1.
The relationship of radicals and incidentals with families and items.

2.1.2. Second-Level Model The item parameters, denoted as $\boldsymbol{\xi}_{i_f}$, are transformed as

$$\boldsymbol{\xi}_{i_f} = (a_{i_f}, b_{i_f}, \text{logit}\, \gamma_{i_f}). \tag{2}$$

We will use c_{i_f} to denote the transformed guessing parameter. Because of this transformation, it can be assumed that the parameters $\boldsymbol{\xi}_{i_f}$ have a multivariate normal distribution

$$\boldsymbol{\xi}_{i_f} \sim \text{MVN}(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f) \tag{3}$$

with $\boldsymbol{\mu}_f$ a vector of mean values for the item parameters and $\boldsymbol{\Sigma}_f$ the covariance matrix of the item parameters for family f.

As an alternative, when the covariance matrices can be assumed to be approximately equal across families, a common covariance matrix can be used,

$$\boldsymbol{\xi}_{i_f} \sim \text{MVN}(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}). \tag{4}$$

The model with family-specific covariance matrices will be labeled LICM-F; the model with a common covariance matrix will be labeled LICM-C. In both models, the mean difficulty of a family is postulated to be a linear combination of the effects of the radicals used to generate an item:

$$\mu_{b_f} = \sum_{r=1}^{R} d_{fr}\beta_r, \tag{5}$$

where β_r is the effect of radical r on the mean difficulty of the item families and d_{fr} is a design variable denoting how often radical r should be used within an item to generate an item from family f. Thus, at the item level,

$$b_{i_f} = \sum_{r=1}^{R} d_{fr}\beta_r + \varepsilon_{i_f}, \quad \varepsilon_{i_f} \sim N\big(0, \sigma^2_{b_f}\big), \tag{6}$$

with $\sigma^2_{b_f}$ the second diagonal element of $\boldsymbol{\Sigma}_f$. As can be seen from (6), the radicals determine the mean family difficulty parameter μ_{b_f} whereas the incidentals determine the family covariance matrix $\boldsymbol{\Sigma}_f$.

It is assumed that θ has a normal distribution with mean μ_θ and standard deviation σ_θ. We set $\mu_\theta = 0$, and $\sigma_\theta = 1$ to identify the model.

2.2. *Parameter Estimation*

In the studies reported below, the parameters of the model were estimated in a Bayesian framework with data-augmented Gibbs sampling. The specific Gibbs sampling algorithm is described in the Appendix and was programmed in the software environment R (R Development Core Team, 2009).

Independent priors were used for the hyperparameters $\boldsymbol{\lambda} = (\boldsymbol{\mu}_a, \boldsymbol{\beta}, \boldsymbol{\mu}_c)$ and $\boldsymbol{\Sigma}_f$ (LICM-F) or $\boldsymbol{\Sigma}$ (LICM-C). A convenient prior for $\boldsymbol{\lambda}$ is the multivariate normal distribution with mean $\boldsymbol{\lambda}_0$ and covariance matrix $\boldsymbol{V}_0$,

$$\boldsymbol{\lambda} \sim \text{MVN}(\boldsymbol{\lambda}_0, \boldsymbol{V}_0). \tag{7}$$

The prior for $\boldsymbol{\Sigma}_f$ was the inverse-Wishart distribution with sum of squares $\boldsymbol{S}_0$ and ν_0 as degrees of freedom,

$$\boldsymbol{\Sigma}_f \sim \text{inverse-Wishart}(\boldsymbol{S}_0, \nu_0). \tag{8}$$

For this prior density to have a finite integral ν_0 should not be smaller than the dimension of $\boldsymbol{\Sigma}_f$ (Gelman, Carlin, Stern, & Rubin, 2004).

Let $\boldsymbol{U} = ((U_{ifn}))$, $\boldsymbol{\xi} = (\boldsymbol{\xi}_f) = ((\boldsymbol{\xi}_{if}))$; the other boldfaced parameters are defined analogously. Combining the above priors with the likelihood results in the following joint posterior for the LICM-F:

$$\begin{aligned} &p(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda} | \boldsymbol{U}, \boldsymbol{Q}) \\ &\quad \propto \prod_{f=1}^{F} \left[p(\boldsymbol{Z}, \boldsymbol{W} | \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\xi}_f) p(\boldsymbol{\xi}_f | \boldsymbol{\lambda}, \boldsymbol{Q}_f, \boldsymbol{\Sigma}_f) p(\boldsymbol{\Sigma}_f | \boldsymbol{S}_0, \nu_0) \right] p(\boldsymbol{\theta}) p(\boldsymbol{\lambda} | \boldsymbol{\lambda}_0, \boldsymbol{V}_0), \end{aligned} \tag{9}$$

in which $\boldsymbol{Q}_f$ is a design matrix such that $\boldsymbol{\mu}_f = \boldsymbol{Q}_f \boldsymbol{\lambda}$. The composition of $\boldsymbol{Q}_f$ and the data-augmentation variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are explained in the Appendix. The posterior density of LICM-C is obtained by replacing $\boldsymbol{\Sigma}_f$ in (9) by $\boldsymbol{\Sigma}$.

Results from a Markov chain can be used as draws from the full posterior only upon convergence of the chain. In the literature, several convergence diagnostics have been proposed. However, none of them is foolproof, and the use of multiple diagnostics to assess different aspects of the convergence is generally recommended. In the studies below, we used Geweke's (1992), Heidelberger and Welch's (1983) and Raftery and Lewis' (1992) diagnostics to assess convergence. Geweke's (1992) diagnostic is a Z-score test for the equality of the means of the first 10% and last 50% of the values drawn in the Markov chain after the burn-in period. Heidelberger and Welch's (1983) and Raftery and Lewis' (1992) diagnostics are based on a criterion of accuracy for the estimated mean and a quantile q of the parameter distributions, respectively. All convergence diagnostics are available in the package Coda for R (Plummer, Best, Cowles, & Vines, 2006).

3. Empirical Study

The example is an analysis of the Analogies subtest of the SON-R 5 1/2-17 non-verbal intelligence test (Laros & Tellegen, 1991; Tellegen & Laros, 1993). Each item of the subtest consisted of three different pictures composed of geometrical figures (A, B, and C). The test taker had to choose a fourth picture (D) from a set of four alternatives such that the transformation(s) applied to C to obtain D (such as form changes and rotations) were the same as those applied to A to create B; that is, the test taker had to complete A : B = C : ?. An example item similar to the items in the Analogies subtest of the SON-R 5 1/2-17 is presented in Figure 2. The test was taken by 1,350 children of age 6–14.

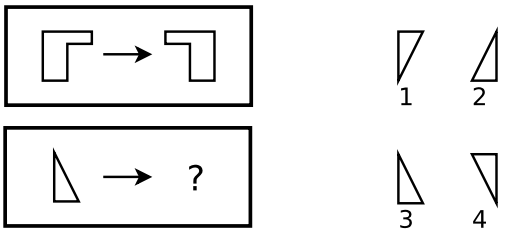

FIGURE 2.
Example item similar to the items in the Analogies subtest of the SON-R 5 1/2-17.

The authors of the test constructed its items by systematically varying their features in accordance with a postulated theory of item difficulty. In all, they distinguished 11 different levels of difficulty and created three items at each level. We interpret this as 11 families with three items each. The difficulty levels were used in an adaptive administration of the test, which ran as follows: The items were combined into three series, where each series contained one item from every family. Within a series, the 11 families were ordered from easy to difficult. Every test taker started with the first item in the first series, and continued with the series until two incorrect answers were given. He or she then continued with item m in the second series, where m was one less than the number of items scored correctly in the first series. The procedure continued similarly, with missing responses at the start of the second series counted as correct responses in the computation of m, until two incorrect answers were given to the items in the third series or the last item in this series had been reached. Because of the adaptive nature of the procedure, items too difficult for a particular test taker were not administered.

As guessing was expected not to be an issue, the 11 item families were analyzed using the 2PNO model as the first-level model. We nevertheless investigated possible item misfit due to guessing by means of a Bayesian latent residual analysis (see Section 3.3). As the number of items per family was low, we used the LICM-C. Missing data due to the adaptive design of the test, both at the beginning and the end of the series, can be treated as missing at random. This is justified because the item administration design was completely determined by the observed responses. Because of this, the ignorability principle for missing data (Rubin, 1976; Glas, 2010) holds, and bias in the parameter estimates was avoided.

The total number of observations per item ranged from 139 to 1,350; per family the range was from 680 (Family 11) to 2,483 (Family 5). The prior means of the family discrimination and effect parameters $\boldsymbol{\lambda}_0$ were set equal to 1. The variances of the prior scale matrix $\boldsymbol{S}_0$ were set equal to 0.1 and the covariances to 0.05. Furthermore, $\boldsymbol{V}_0$ was set equal to a diagonal matrix with elements 100 (representing a case of low prior information) and ν_0 was set equal to 2 (i.e., smallest value of ν_0 resulting in a prior distribution with a finite integral; Gelman et al., 2004).

Expected a posteriori (EAP) estimates of the hyperparameters of the LICM-C were computed from 100,000 iterations of the Gibbs sampler after the first 20,000 iterations for burn-in. Convergence of the sampler for the hyperparameters was checked using Geweke's (1992) and Heidelberger and Welch's (1983) diagnostics and by inspecting convergence plots.

3.1. Item Structures

Laros and Tellegen (1991) created the items according to the theory that item difficulty increases with (1) the number of transformations performed on the A-term, (2) the number of basic elements in the A-term, (3) the complexity of the transformations on the A-term, (4) the dissimilarity between the A-term and the C-term, and (5) the similarity between the correct and incorrect alternatives. For example, to solve the item in Figure 2 only one transformation on one basic element is needed (the transformation is to mirror the triangle on the vertical axis). As the

$\boldsymbol{D}_1$

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

FIGURE 3.
Design matrix of the baseline model $\boldsymbol{D}_1$.

item families were not systematically designed with respect to the other three factors (see Laros & Tellegen, 1991; Tellegen & Laros, 1993), we tested the hypothesis that the first two rules explained the family difficulty parameters against the alternative that the effects of these other factors could not be ignored.

To test the hypotheses, three models with different design matrices were constructed. The first model contained a different dummy rule for each family. The design matrix for this baseline model, $\boldsymbol{D}_1$, is given in Figure 3. The rows of the matrix correspond to the different families and the columns to the different rules, except for the first column, which introduces the first family as a reference family. Note that fitting the LICM for this design matrix is comparable to fitting the ICM by Glas and van der Linden (2003). This is the most general model we will consider.

The other two models were constructed to test the effect of the two rules. For the first ten families, all three items within each family had the same value for the two rules. Family 11 did not systematically vary according to these rules. Therefore, this family was not modeled by the two rules, but by a family-specific intercept. In this way, all 11 families could be analyzed in the same run. As noted above, removing the items from Family 11 from the analysis would have caused a violation of the ignorability principle.

The design matrices for the first ten families were constructed to investigate the effect of the number of transformations and the number of basic elements in the items (see D_2 and D_3 in Figure 4). In both matrices, the entries in the second and third column represent the effect of adding a specific number of transformations and basic elements, respectively, to the item. In model D_2, five additional columns were added to account for differences in family difficulty due to other difficulty factors which might be present in the items.

The design matrix of model D_3 only contained the two rules for the number of transformations and the number of basic elements. Observe that in this model Families 1–4, 5 and 7, and 8–9 were restricted to have the same mean difficulty. However, the means of their discrimination parameters were allowed to vary.

$$\boldsymbol{D}_2 \qquad\qquad \boldsymbol{D}_3$$

$$\begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\
1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 2 & 0 & 0 & 0 & 1 & 0 \\
1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\
1 & 2 & 2 & 0 & 0 & 0 & 0 & 1 \\
1 & 2 & 3 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 2 \\
1 & 2 & 1 \\
1 & 1 & 2 \\
1 & 2 & 2 \\
1 & 2 & 2 \\
1 & 2 & 3
\end{bmatrix}$$

FIGURE 4.
Design matrices of the two restricted models $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$.

3.2. Model Comparison and Model Fit

To test the hypotheses mentioned above, the three models were compared by means of the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). The DIC is a model selection criterion based on a measure of model fit, $\bar{D}(\boldsymbol{\eta})$, and a penalty for model complexity, p_D:

$$\text{DIC} = \bar{D}(\boldsymbol{\eta}) + p_D, \tag{10}$$

where $\boldsymbol{\eta}$ are the parameters of the model. Define the deviance as -2 times the log likelihood:

$$D(\boldsymbol{\eta}) = -2 \log \prod_{f=1}^{F} p(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{\xi}_f) p(\boldsymbol{\xi}_f|\boldsymbol{\mu}_f, \boldsymbol{\Sigma}). \tag{11}$$

Model fit, $\bar{D}(\boldsymbol{\eta})$, is then defined as the posterior mean of the deviance, and model complexity, p_D, as $\bar{D}(\boldsymbol{\eta})$ minus the deviance at the posterior mean of the parameters $D(\bar{\boldsymbol{\eta}})$. $\bar{D}(\boldsymbol{\eta})$ and $D(\bar{\boldsymbol{\eta}})$ can be estimated using posterior simulations of the parameters. The use of the DIC is such that the model with the smallest value for the DIC is to be preferred.

For the LICM, it is important to investigate whether the radicals can properly explain the family difficulty parameters. There are several reasons for a possible discrepancy between the family difficulty parameters in the ICM and LICM. First of all, the set of radicals in the LICM may be incomplete. Similarly, the design matrix may have been misspecified. For example, omitting a term for an interaction between certain radicals that may have occurred may result in bias in the estimated family difficulty parameters. Finally, the assumption of a linear relationship between the radicals and the family difficulty parameters may not hold.

To investigate whether the specified radicals and design matrix could properly explain the family difficulty parameters, a statistic based on a suggestion by one of the reviewers was applied. For each family, the mean and 95% highest posterior density (HPD) interval of the difference between the empirical and modeled means of the item parameters were computed across iterations of the Gibbs sampler. A HPD interval for this statistic not including zero was taken as a sign

of model misspecification. Conclusions based on the statistic may be conservative, because the family mean parameters also serve as means of the prior distributions in the estimation of the item parameters. However, the impact of the means of the family priors is moderated by the estimated covariance matrix, which will be larger when an estimated family difficulty parameter lies further away from the empirical mean of the item difficulty parameters. The impact can therefore be expected to automatically decrease with increasing bias in the family parameters.

An alternative way of identifying misfit due to the linear structure on the family difficulty parameters is to compare the ICM and the LICM with regard to the family difficulty estimates, the explained variance at the level of the item parameters, and the degree of pooling of the item parameters around their family means. To this end, Gelman and Pardoe's (2006) explained variance, R^2, and pooling factor, λ, were computed. (The standard notation for the pooling factor should not be confused with that for the hyperparameters of the model.) The explained variance in the item difficulty parameters can be computed as

$$R_b^2 = 1 - \frac{E[\text{Var}(b_{i_f} - \mu_{b_f})]}{E[\text{Var}(b_{i_f})]}, \tag{12}$$

where E represents the mean over the posterior simulations, and Var represents the finite-sample variance operator over the parameters. The pooling factor of the item difficulty parameters can be computed as

$$\lambda_b = 1 - \frac{\text{Var}[E(b_{i_f} - \mu_{b_f})]}{E[\text{Var}(b_{i_f} - \mu_{b_f})]}. \tag{13}$$

A pooling factor of less than 0.5 indicates a higher degree of within-family than between-family information. The explained variance and pooling factor for the discrimination parameters, R_a^2 and λ_a, can be computed analogously.

Bayesian latent residual analyses were performed to further investigate the fit of the three models to the data (Fox, 2004; Johnson & Albert, 1999). For the 2PNO, the Bayesian latent residual corresponding to response $U_{i_f n}$ can be defined as

$$\epsilon_{i_f n} = Z_{i_f n} - a_{i_f}\theta_n + b_{i_f}, \tag{14}$$

where $Z_{i_f n}$ is a data-augmentation variable explained in the Appendix. Outliers were defined as observed responses with absolute residuals greater than two standard deviations. The posterior probabilities of correct or incorrect responses being outliers were computed from the draws of the Markov chain as

$$\begin{aligned} p\big(|\epsilon_{i_f n}| > 2 | U_{i_f n} = 1, \theta_n, \boldsymbol{\delta}_{i_f}\big) &= \frac{\Phi(-2)}{\Phi(a_{i_f}\theta_n - b_{i_f})}, \\ p\big(|\epsilon_{i_f n}| > 2 | U_{i_f n} = 0, \theta_n, \boldsymbol{\delta}_{i_f}\big) &= \frac{\Phi(-2)}{1 - \Phi(a_{i_f}\theta_n - b_{i_f})}, \end{aligned} \tag{15}$$

respectively. A high percentage of outlying responses indicates a poor model fit. In particular, many outliers among the correct responses can be an indication that guessing occurred.

3.3. Results

Table 1 presents the values of the DIC and its constituent model fit measure $\bar{D}(\boldsymbol{\eta})$, model complexity measure p_D, the explained variance R^2 and pooling factor λ for the three models. As indicated by its value for $\bar{D}(\boldsymbol{\eta})$, model $\boldsymbol{D}_2$ did fit the data almost as well as the most general model $\boldsymbol{D}_1$. Ignoring the residual variance in the family parameters (model $\boldsymbol{D}_3$) did lead to a more

TABLE 1.
Summary statistics (DIC, $\bar{D}(\boldsymbol{\eta})$, p_D, R^2, and λ) for the three models.

Model	DIC	$\bar{D}(\boldsymbol{\eta})$	p_D	R_a^2	λ_a	R_b^2	λ_b
$\boldsymbol{D}_1$	22544	21332	1212	0.581	0.529	0.905	0.439
$\boldsymbol{D}_2$	22549	21337	1212	0.509	0.523	0.876	0.364
$\boldsymbol{D}_3$	22562	21355	1207	0.426	0.543	0.775	0.177

TABLE 2.
Expected a posteriori estimates and 95% highest posterior density intervals of the radical and (co)variance parameters.

Radical	$\boldsymbol{D}_1$		$\boldsymbol{D}_2$		$\boldsymbol{D}_3$	
	EAP($\boldsymbol{\beta}_r$)	HPD($\boldsymbol{\beta}_r$)	EAP($\boldsymbol{\beta}_r$)	HPD($\boldsymbol{\beta}_r$)	EAP($\boldsymbol{\beta}_r$)	HPD($\boldsymbol{\beta}_r$)
No. transf.	–	–	1.249	[0.783, 1.707]	0.907	[0.451, 1.368]
No. elem.	–	–	0.747	[0.435, 1.057]	0.941	[0.606, 1.293]
(Co) Variance	EAP($\boldsymbol{\Sigma}$)	HPD($\boldsymbol{\Sigma}$)	EAP($\boldsymbol{\Sigma}$)	HPD($\boldsymbol{\Sigma}$)	EAP($\boldsymbol{\Sigma}$)	HPD($\boldsymbol{\Sigma}$)
σ_a^2	0.061	[0.020, 0.116]	0.068	[0.021, 0.127]	0.081	[0.024, 0.155]
σ_{ab}	−0.013	[−0.074, 0.038]	−0.016	[−0.089, 0.054]	−0.029	[−0.164, 0.097]
σ_b^2	0.124	[0.045, 0.227]	0.156	[0.066, 0.269]	0.284	[0.133, 0.465]

parsimonious model. However, the decrease in model complexity did not compensate the value of the DIC for the increase in model misfit.

As expected, decreasing the number of second-level parameters in the model resulted in less variance explained in the item parameters according to R_b^2 for the three models. Also, when fewer parameters were present in the model to explain the family difficulty parameters, the item difficulty parameters were less pooled around their family means, as indicated by λ_b. This was also reflected by the larger estimates of the within-family variance of the difficulty parameters (see Table 2) for the more restrictive models. Especially for model $\boldsymbol{D}_3$, the differences in R_b^2, λ_b, and σ_b^2 with the baseline model, $\boldsymbol{D}_1$, were large.

Tables 2, 3, and 4 present the EAP estimates and the 95% HPD intervals for the parameters for the radicals, the common covariance matrices, and the family means for the discrimination and difficulty parameters for the three models. The two radicals of interest in models $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$ had a large positive effect on the difficulty of the items. Thus, both an increase of the number of transformations ("No. transf." in Table 2) needed to solve the item and the number of basic elements ("No. elem." in Table 2) in it reduced the probability of a correct answer. For the model with family-specific intercepts to account for unexplained variance in the mean family difficulty, $\boldsymbol{D}_2$, the effect of the number of transformations was larger than the effect of the number of basic elements. For the model without a distinction between families with the same combination of radicals, $\boldsymbol{D}_3$, the difference was reversed.

In all three models, the estimate of the common family covariance matrix revealed negative covariance between the discrimination and difficulty parameters. Both within (Table 2) and between the families (Tables 3 and 4), the easier items tended to have larger discrimination parameters than the more difficult items.

The mean discrimination per family was very similar across the three models (see Table 3). However, the mean family difficulty showed some variation (see Table 4), with the parameters for model $\boldsymbol{D}_2$ being more similar to those of the baseline model $\boldsymbol{D}_1$ than the parameters for model $\boldsymbol{D}_3$ (as was already indicated by R_b^2). To get an indication as to which families were especially biased by the linear approximation in model $\boldsymbol{D}_3$, we checked which ICM estimates of the family

TABLE 3.
Expected a posteriori estimates and 95% highest posterior density intervals of the family discrimination parameters.

Family	$\boldsymbol{D}_1$		$\boldsymbol{D}_2$		$\boldsymbol{D}_3$	
	EAP($\boldsymbol{\mu}_a$)	HPD($\boldsymbol{\mu}_a$)	EAP($\boldsymbol{\mu}_a$)	HPD($\boldsymbol{\mu}_a$)	EAP($\boldsymbol{\mu}_a$)	HPD($\boldsymbol{\mu}_a$)
1	1.177	[0.796, 1.567]	1.007	[0.631, 1.400]	0.982	[0.574, 1.414]
2	1.230	[0.871, 1.612]	1.236	[0.859, 1.617]	1.188	[0.824, 1.574]
3	1.369	[1.027, 1.716]	1.366	[1.008, 1.729]	1.341	[0.987, 1.714]
4	0.949	[0.628, 1.269]	0.948	[0.612, 1.280]	1.012	[0.621, 1.415]
5	1.084	[0.755, 1.410]	1.118	[0.770, 1.474]	1.076	[0.736, 1.418]
6	0.726	[0.420, 1.040]	0.746	[0.422, 1.078]	0.764	[0.402, 1.136]
7	0.803	[0.489, 1.122]	0.808	[0.481, 1.137]	0.869	[0.406, 1.323]
8	0.685	[0.365, 1.009]	0.695	[0.368, 1.032]	0.693	[0.344, 1.044]
9	0.595	[0.273, 0.909]	0.601	[0.265, 0.934]	0.611	[0.270, 0.954]
10	0.655	[0.308, 1.003]	0.684	[0.320, 1.057]	0.689	[0.283, 1.104]
11	0.374	[0.039, 0.724]	0.391	[0.025, 0.747]	0.388	[0.020, 0.779]

TABLE 4.
Expected a posteriori estimates and 95% highest posterior density intervals of the family difficulty parameters.

Family	$\boldsymbol{D}_1$		$\boldsymbol{D}_2$		$\boldsymbol{D}_3$	
	EAP($\boldsymbol{\mu}_b$)	HPD($\boldsymbol{\mu}_b$)	EAP($\boldsymbol{\mu}_b$)	HPD($\boldsymbol{\mu}_b$)	EAP($\boldsymbol{\mu}_b$)	HPD($\boldsymbol{\mu}_b$)
1	−2.122	[−2.678, −1.572]	−1.632	[−2.050, −1.232]	−1.431	[−1.737, −1.138]
2	−1.850	[−2.372, −1.341]	−1.862	[−2.420, −1.328]	−1.431	[−1.737, −1.138]
3	−1.700	[−2.168, −1.262]	−1.699	[−2.193, −1.188]	−1.431	[−1.737, −1.138]
4	−0.911	[−1.336, −0.496]	−0.910	[−1.376, −0.442]	−1.431	[−1.737, −1.138]
5	−0.603	[−1.016, −0.184]	−0.885	[−1.275, −0.509]	−0.489	[−0.839, −0.142]
6	−0.177	[−0.585, 0.234]	−0.384	[−0.783, 0.016]	−0.524	[−0.980, −0.064]
7	0.258	[−0.158, 0.669]	0.259	[−0.194, 0.723]	−0.489	[−0.839, −0.142]
8	0.222	[−0.188, 0.650]	0.364	[0.081, 0.641]	0.418	[0.109, 0.740]
9	0.534	[0.097, 0.949]	0.542	[0.072, 1.015]	0.418	[0.109, 0.740]
10	0.993	[0.538, 1.467]	1.111	[0.677, 1.552]	1.359	[0.889, 1.843]
11	0.793	[0.312, 1.275]	0.815	[0.286, 1.341]	0.815	[0.159, 1.492]

difficulty parameters were not included in the 95% HPD interval of the respective LICM estimates. This comparison showed estimates for Families 1, 2, 4, and 7 that were especially biased by the linear approximation. This finding was also corroborated by the 95% HPD intervals of the differences between the empirical and modeled family means of the item difficulty parameters. For the same four families, the HPD intervals did not include zero. Based on these results, the hypothesis of two radicals in model $\boldsymbol{D}_3$ (the number of transformations and the number of basic elements in an item) explaining the family difficulty parameters was rejected. Model $\boldsymbol{D}_2$, with its additional parameters to explain the residual variance, provided a better approximation to the family difficulty parameters of the ICM; only the ICM difficulty estimate of the first family was not included in the HPD interval for the LICM estimate but all HPD intervals for the differences between the empirical and family means of the item parameters did include zero. Because this model also had a lower DIC value than the model without the residual variance $\boldsymbol{D}_3$, it was selected and investigated for its fit to the data by means of a Bayesian latent residual analysis. (The model fit analyses for the other two models yielded almost the same results, due to the similarity of their first-level parameters.)

Figures 5 (Items 1 to 15), 6 (Items 16 to 30), and 7 (Items 30 to 33) show the results of this model fit analysis. For each item, the probabilities of the responses being outliers (y-axis),

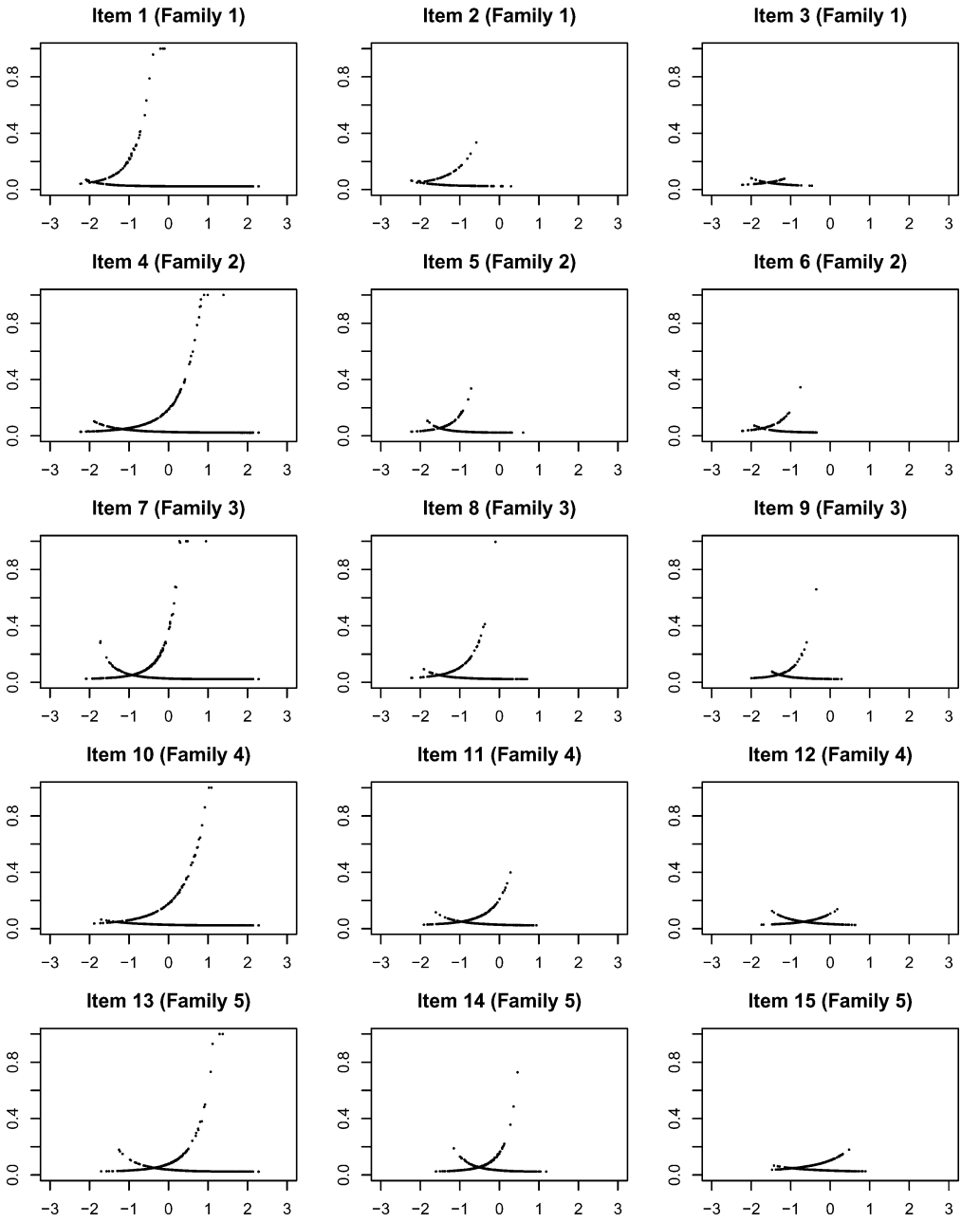

FIGURE 5.
Outlying probabilities as a function of the estimated ability parameter for the items in the first five families. The upward trends relate to the incorrect responses; the downward trends relate to the correct responses.

as computed by (15), were plotted against the EAP estimated ability parameters for the persons giving the responses (x-axis). Of the 20,863 responses, the probabilities of some 0.0007% were computed as greater than 1. In Figure 5, these probabilities were set equal to 1.

The outlying probabilities for the correct and the incorrect responses were plotted in the same figures. The curves can be distinguished by their form; the upward trends relate to the

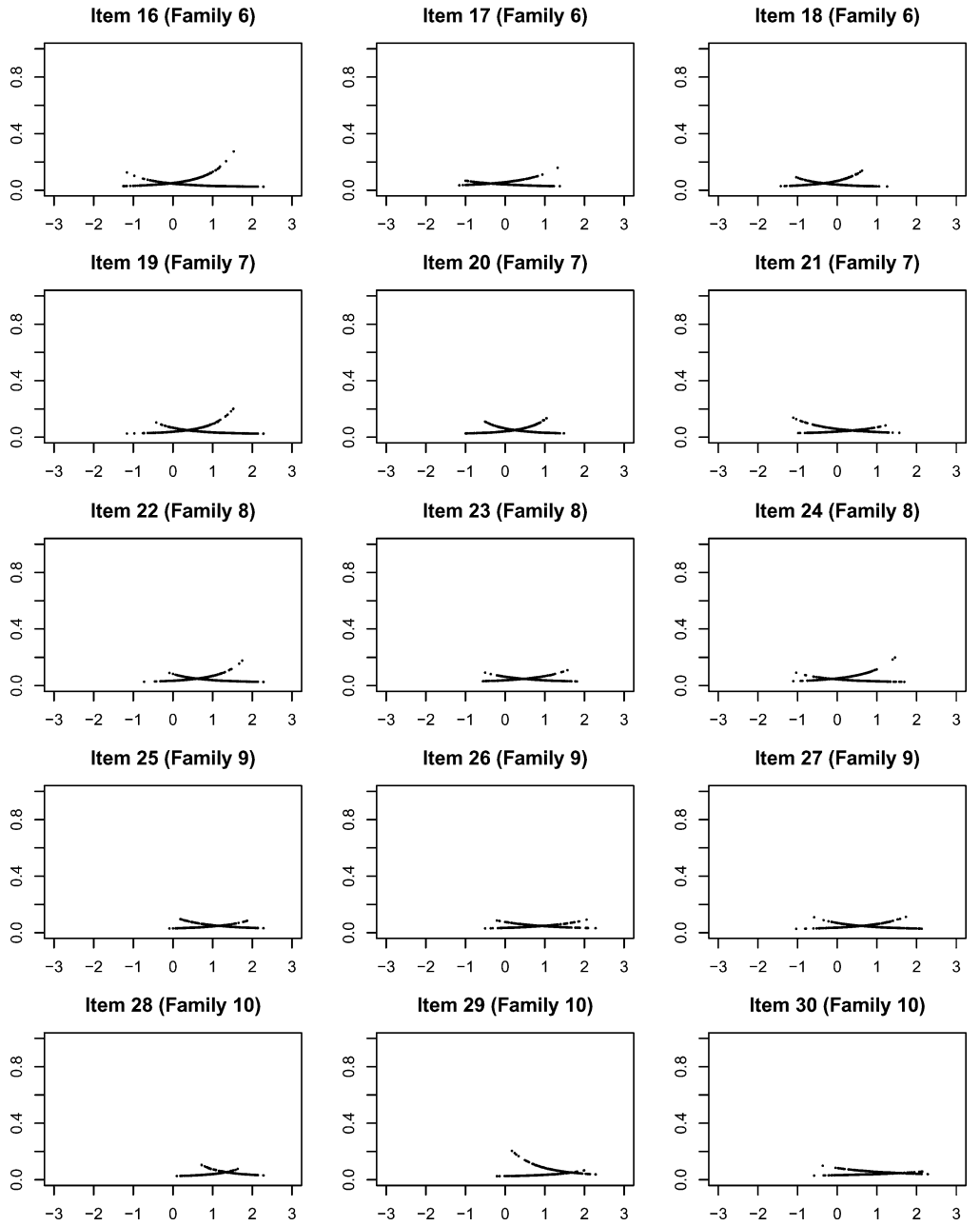

FIGURE 6.
Outlying probabilities as a function of the estimated ability parameter for the items in Family 6 to 10. The upward trends relate to the incorrect responses; the downward trends relate to the correct responses.

incorrect responses, whereas the downward trends relate to the correct responses. The former can be explained as follows. When a person of high ability gives an incorrect response to an easy item, the probability of the response being an outlier is larger than that of a person of low ability giving an incorrect response to the same item. The downward trends for the correct responses are explained analogously.

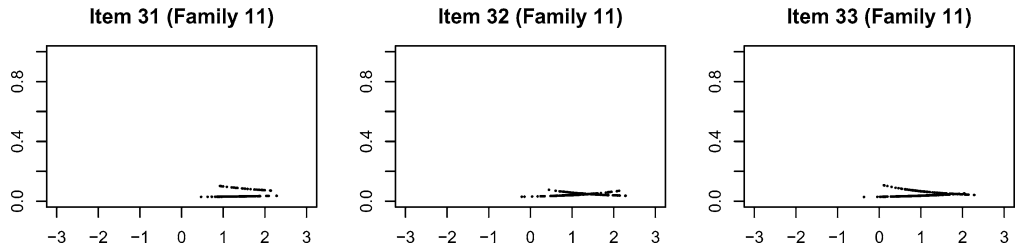

FIGURE 7.
Outlying probabilities as a function of the estimated ability parameter for the items in Family 11. The upward trends relate to the incorrect responses; the downward trends relate to the correct responses.

In our model fit analysis, we did not find any indication of model misfit due to guessing. If guessing had occurred, large outlying probabilities for the *correct* responses should have been observed. However, only five of the items yielded large probabilities for the *incorrect* responses (Items 1, 4, 7, 10, and 13). These items were the first items from the first five families. For most test takers, they were the first items met in the adaptive test. Therefore, it seems likely that they needed to warm up to the task—a process that might have resulted in more incorrect answers than expected given their ability level.

Note also the effect of the adaptive method of item administration in the plots. The first few items in the first series (Items 1, 4, 7, etc.) were answered by almost all test takers and, consequently, a relatively large range of estimated abilities on the x-axis was observed. The items in the second (Items 2, 5, 8, etc.) and third (Items 3, 6, 9, etc.) series were only answered by test takers with an estimated ability at the location of the difficulty of the item. For example, Item 3 (Family 1) was only answered by test takers with a negative ability estimate, whereas most of the test takers answering Item 33 (Family 11) had a positive ability estimate.

3.4. Conclusion

By comparing the fit of LICM models of different complexity, hypotheses with regard to the radicals can be tested. Based on such analyses, the set of radicals can be adapted until the selected combination of radicals satisfactorily explains the family difficulty parameters, and extra parameters to account for residual variance are unnecessary to obtain a good model fit. Such a set of calibrated radicals facilitates item generation on-the-fly. For instance, in adaptive testing, based on a test takers' updated ability estimate, a family can be optimally selected (see Glas & van der Linden, 2003), and a new item can be generated according to the design of the selected family as specified in the design matrix. Subsequently, after an answer has been given to the item, the ability estimate of the test taker can be updated using the estimates of the family distributions $\boldsymbol{\mu}_f$ and $\boldsymbol{\Sigma}_f$.

4. Simulation Study

A parameter recovery study was conducted to explore the impact of the sampling design on the calibration of the second-level parameters in the model. In this study, each design meets the requirement that every test taker responds to one item from every family. The research question is: Conditional on the number of test takers, how should we sample the items? More specifically, if our goal is to estimate the second-level parameters as accurately as possible, would it be better to sample:

- more items per family, reducing the numbers of test takers per item, or
- sample fewer items per family, thereby increasing the numbers of test takers per item?

4.1. Study Setup

The design of this study consisted of runs with eight or 16 families and 10 or 20 items per family. The total number of test takers was fixed at either 500 or 1000. Each cell of the design was replicated 20 times. The data were simulated and the parameters were estimated under the assumption of equal covariance matrices across families (LICM-C) with the 2PNO as the first-level model. The choice for the 2PNO was made for practical reasons only. In the Appendix it is explained that the conditional posterior distribution of the guessing parameters is not a standard distribution, and therefore a method such as importance sampling would have been needed to sample for these parameters. Consequently, more iterations would have been required for the Gibbs sampler to converge. Because of the large number of runs in our study, the total running time would have been prohibitive, while the feasibility of the method for sampling γ_{i_f} has already been demonstrated by Glas et al. (2010).

As the simulated test takers were required to respond to one item from every family, the number of observations per item N_{i_f} was equal to the total number of test takers N_f divided by the number of items per family I_f. Parameters $\boldsymbol{\beta}$ were set equal to 2.9, 1.1, 0.8, so that the application of a rule had a positive effect on the difficulty of the items; the common effect was chosen to be -2.7.

For the case of eight families, a full factorial design was used:

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

For the case of 16 families, the rows in the matrix were duplicated. This was done to avoid confounding of the number of families with the structure of the design matrix when comparing the precision of parameter estimates. Without duplication of the design matrix, adding new families would have required the introduction of more radicals, and would have resulted in a design matrix with a different structure for the conditions with 16 families than the design matrix for the conditions with eight families. Note that, in the estimation procedure, families with the same combination of radicals will have the same estimated family difficulty parameter, but their estimated discrimination parameters are allowed to vary.

The family discrimination parameters were fixed at one and the difficulty parameters at $\boldsymbol{D\beta}$. The covariance matrices were chosen to be equal across families (LICM-C); the diagonal of the common matrix was equal to (0.10, 0.10) and the off-diagonal elements were equal to 0.05.

In every replication, the item parameters were randomly drawn from their family distributions, according to (4). Negative draws for the discrimination parameters were discarded. The ability parameters were drawn from the standard normal distribution.

The prior means of the family discrimination and effect parameters $\boldsymbol{\lambda}_0$ as well as the prior scale matrix $\boldsymbol{S}_0$ were set equal to their true values. Furthermore, $\boldsymbol{V}_0$ was set equal to the $F + R$ identity matrix and ν_0 was set equal to 2 (i.e., smallest value of ν_0 resulting in a prior distribution with a finite integral; Gelman et al., 2004).

The convergence diagnostics of Geweke (1992), Heidelberger and Welch (1983), and Raftery and Lewis (1992) were used to determine the total number of iterations as well as the number of burn-in iterations needed for the Gibbs sampler to reach satisfactory convergence. To

reduce the total running time for the study, we used the true values of the parameters as starting values to create a smaller number of burn-in iterations—a choice that did not influence the final results. In every replication, 32,000 iterations were performed, 2,000 of which had to be discarded as a burn-in. Following a recommendation by MacEachern and Berliner (1994), the Gibbs sampler was not subsampled and all iterations after burn-in were used to compute both the EAP estimates of the parameters as well as their mean absolute error (MAE), bias, and variance.

4.2. Results

The first three columns of Table 5 give the number of families F, items per family I_f, and persons N_f in the design, respectively. Also, the resulting number of observations per item N_{i_f} is given. The next three columns give the MAEs for the first-level parameter estimates, and the last columns give the MAEs for the second-level parameter estimates.

Because the values for the MAE depend on the scale that was set for the model, the content of the table should be interpreted in a relative manner; that is, to compare between the conditions. However, to give some feeling for the size of the MAEs, the results can be compared to the results for the two-parameter logistic (2PL) model reported in van der Linden and Glas (2000). These authors also fixed the scale by setting $\boldsymbol{\theta} \sim N(0, 1)$. With a test length of 20 items and a true ability value of zero, they found a MAE for the estimated ability parameters of approximately 0.3 when using a weighted maximum likelihood estimation procedure and a maximum-information criterion for selecting the items from an item pool consisting of 400 items (van der Linden & Glas, 2000; Figure 5). In our simulation study, we found a similar value of approximately 0.35 with a test length of 16 items and an item pool of 320 items (Table 5; sixth row).

In general, the higher-level parameters were better recovered than the first-level parameters. Because every simulated test taker responded to one item from every family, the conditions with 16 families resulted in a lower MAE for the ability parameters and, consequently, in lower MAEs for all other parameters, as compared to the conditions with only eight families.

In addition, the number of families had a larger effect on the MAE of the estimated family difficulty parameters than for the estimated family discrimination parameters. This was as expected. Across families, every item with a common effect parameter β_r contributes to the accuracy of the estimate of this parameter and, consequently, to the accuracy of the family difficulty estimates computed from the estimate. In contrast, only items in the same family contribute to the estimate of its discrimination parameter.

Generally, the MAEs for the covariance matrix were small. This finding is assumed to be due to the assumption of a common matrix across families: Under the assumption, the estimate of the covariance matrix is based on $F * I_f$ observations—a product ranging from 80 to 320 in this study. When family-specific covariance matrices are to be estimated (i.e., when the LICM-F is used), only I_f observations are available to estimate each covariance matrix.

To answer our earlier research question, for each level F, the rows of the table with equal N_f should be compared. Generally, a larger number of observations per item resulted in lower MAEs for the item parameters. However, for the second-level parameters, lower MAEs were obtained for larger numbers of items per family, even though this resulted in lower numbers of observations per item.

Tables 6 and 7 present the results in terms of mean bias and mean variance. In a simulation study with a limited number of replications, it is important to check whether the results could be due to chance. Therefore, for all types of parameters, t-tests for independent samples were conducted on the mean bias and the mean variances to check whether their differences between conditions with an equal number of families F and an equal number of test takers N_f were significant over the 20 replications. At a level of 0.05, the t-tests on the mean bias did not reveal any significant differences between the conditions. In all conditions, the bias was small, and differences between the conditions should be due to chance. However, the t-tests on the mean

TABLE 5.
Mean absolute error of the estimated first-level and second-level parameters as a function of the number of families F, items within families I_f and persons N_f.

					Item par.		Family par.			Cov. matrix		
F	I_f	N_f	N_{i_f}	$\boldsymbol{\theta}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{\mu}_a$	$\boldsymbol{\mu}_b$	$\boldsymbol{\beta}$	$\boldsymbol{\sigma}_a^2$	$\boldsymbol{\sigma}_{ab}$	$\boldsymbol{\sigma}_b^2$
8	10	500	50	0.445	0.215	0.205	0.145	0.114	0.119	0.026	0.025	0.031
	20	500	25	0.455	0.241	0.220	0.129	0.103	0.118	0.031	0.015	0.024
	10	1000	100	0.444	0.196	0.170	0.116	0.091	0.102	0.024	0.015	0.022
	20	1000	50	0.443	0.214	0.194	0.112	0.083	0.086	0.023	0.014	0.018
16	10	500	50	0.352	0.205	0.188	0.122	0.065	0.056	0.019	0.015	0.016
	20	500	25	0.347	0.226	0.207	0.108	0.063	0.068	0.020	0.015	0.018
	10	1000	100	0.343	0.173	0.159	0.097	0.052	0.055	0.027	0.015	0.016
	20	1000	50	0.341	0.200	0.183	0.079	0.056	0.044	0.021	0.013	0.015

TABLE 6.
Mean bias of the estimated first-level and second-level parameters as a function of the number of families F, items within families I_f and persons N_f.

				Item par.		Family par.			Cov. matrix		
F	I_f	N_f	$\boldsymbol{\theta}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{\mu}_a$	$\boldsymbol{\mu}_b$	$\boldsymbol{\beta}$	$\boldsymbol{\sigma}_a^2$	$\boldsymbol{\sigma}_{ab}$	$\boldsymbol{\sigma}_b^2$
8	10	500	0.012	0.029	0.001	0.034	0.001	0.015	−0.012	0.001	0.013
	20	500	0.015	0.039	0.020	0.038	0.019	0.026	−0.016	0.002	0.010
	10	1000	−0.006	0.021	−0.010	0.018	−0.015	0.005	0.008	−0.001	−0.003
	20	1000	0.002	0.012	0.009	0.019	0.008	0.009	−0.009	0.007	0.009
16	10	500	−0.014	0.033	−0.004	0.027	0.003	0.010	0.002	0.002	0.006
	20	500	0.004	0.019	−0.008	0.016	−0.013	−0.001	−0.007	0.005	0.002
	10	1000	−0.015	0.014	−0.024	0.025	−0.009	0.002	0.007	0.003	−0.003
	20	1000	−0.016	−0.002	−0.018	0.003	−0.010	−0.005	−0.007	−0.001	−0.001

variances did produce significant results (see Table 7). In all comparisons, the family mean parameters $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$, and the effect parameters $\boldsymbol{\beta}$ were estimated with higher precision when the number of items within families was larger.

4.3. Conclusion

A conspicuous advantage of the ICM and the LICM is that they do not require the calibration of newly generated items from families. However, the advantage does require an investment in the form of the calibration of the families (i.e., estimation of the second-level parameters). As shown in the simulation study, careful balancing between the number of items per family and the number of persons to whom an item from it is administered reduces the amount of data to be collected to obtain a certain precision.

5. Discussion

A model for the calibration of items generated by rule-based cloning algorithms was presented. The model is applicable to situations where items within families are generated by the

TABLE 7.
Mean variance of the estimated first-level and second-level parameters as a function of the number of families F, items within families I_f and persons N_f.

				Item par.		Family par.			Cov. matrix		
F	I_f	N_f	$\boldsymbol{\theta}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{\mu}_a$	$\boldsymbol{\mu}_b$	$\boldsymbol{\beta}$	$\boldsymbol{\sigma}_a^2$	$\boldsymbol{\sigma}_{ab}$	$\boldsymbol{\sigma}_b^2$
8	10	500	0.304	0.074	**0.068**	**0.035**	**0.020**	**0.021**	0.002	0.001	0.002
	20	500	0.310	0.081	**0.080**	**0.028**	**0.016**	**0.017**	0.001	0.001	0.001
	10	1000	0.305	0.057	**0.045**	**0.024**	**0.013**	**0.014**	**0.001**	**0.001**	0.001
	20	1000	0.308	0.063	**0.059**	**0.017**	**0.010**	**0.011**	**0.001**	**0.000**	0.001
16	10	500	0.179	**0.068**	**0.058**	**0.027**	**0.010**	**0.009**	**0.001**	**0.000**	0.001
	20	500	0.185	**0.075**	**0.068**	**0.021**	**0.008**	**0.007**	**0.001**	**0.000**	0.001
	10	1000	**0.177**	**0.047**	**0.039**	**0.019**	**0.006**	**0.006**	**0.001**	**0.000**	**0.000**
	20	1000	**0.185**	**0.057**	**0.052**	**0.013**	**0.005**	**0.005**	**0.000**	**0.000**	**0.000**

The boldfaced numbers indicate a significant difference at the 0.05 level.

same rules. These rules ("radicals") are modeled as fixed effects whereas the joint effects of all irrelevant item features ("incidentals") are modeled as random effects. In the current testing practice, it is still somewhat unusual to consider item parameters as random. But from a theoretical viewpoint it makes sense to view them as sampled from families defined by different content rules. For a more complete discussion of the advantages of treating item parameters as random, see De Boeck (2008).

In practical applications, the status of the radicals and incidentals as fixed and random effects should be checked against empirical results. For example, for the earlier example of a statistics item, one might expect the actual numerical information provided in the context story not to influence the difficulty of the item. However, the assumption has to be rejected when some types of numerical information, for example large numbers, lead to substantially higher difficulty estimates than others. In such cases, an assumed incidental should be turned into a radical.

Of course, the success of automated item generation depends on how well the radicals explain the family difficulty parameters. For instance, an assumed effect of a radical may not be found. Also, the possibility of interaction effects between radicals should be considered. Procedures have been presented to investigate the degree of misfit due to unexplained variance in the family difficulty parameters. If the degree of misfit is unacceptable and the cause for misfit can be identified, it may be worth considering an adaptation of the model. Alternatively, it would be interesting to investigate whether the model can be extended with an error term to account for residual variance in the family difficulty parameters.

Also, a basic assumption of the model is that the radicals are assumed to have the same effect for every individual test taker, but this may not always hold. For example, if different strategies for an item lead to the same solution, a radical may be more difficult for a student using one strategy but easier for a student using another. Other models may then provide a better description of the data. In future research, we plan to explore a larger array of model selection and model fit procedures potentially useful for checking both the hierarchical and regression structure of the proposed model.

As for the choice of incidentals, a trade-off exists between the likelihood of generating items with a different appearance and keeping the family covariance matrices small. Smaller covariance matrices are expected to lead to more accurate estimation of the ability parameters. Also, the amount of uncertainty in the item parameters influences the information provided by a test assembled from a pool of item families using optimal test assembly methods.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG), Schwerpunktprogramm "Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293), Project "Rule-based Item Generation of Algebra Word Problems Based upon Linear Logistic Test Models for Item Cloning and Optimal Design." The authors are grateful to Peter Tellegen and Jacob Laros for the permission to use their data on the SON-R 5 1/2-17 in the empirical study.

Appendix: Gibbs Sampling Algorithm

The Gibbs sampler used for the analyses in this paper is based on a reparameterization of the 3PNO model; that is, we assume that the normal distribution function in (1) is replaced by $\Phi(a_{i_f}\theta_n - b_{i_f})$.

A.1. Drawing the First-Level Parameters

Albert (1992) devised a Gibbs sampling scheme to estimate the parameters of the 2PNO model. In order to be able to obtain simple conditional distributions, he augmented the observed data with latent data (Tanner, 1996). An extension to the 3PNO model was suggested in Johnson and Albert (1999, Section 6.9; see also Béguin & Glas, 2001). The suggestion is based on the assumption that person n knows the correct answer to item i_f with probability $\Phi(a_{i_f}\theta_n - b_{i_f})$ and then gives a correct response with probability one. Alternatively, the person does not know the answer with probability $(1 - \Phi(a_{i_f}\theta_n - b_{i_f}))$ and then guesses the correct answer with probability γ_{i_f}. The marginal probability of a correct response is the sum of the probabilities associated with these two processes,

$$p(U_{i_f n} = 1|\theta_n, \boldsymbol{\xi}_{i_f}) = \Phi(a_{i_f}\theta_n - b_{i_f}) + \gamma_{i_f}\big(1 - \Phi(a_{i_f}\theta_n - b_{i_f})\big). \tag{16}$$

This interpretation suggests the introduction of latent augmentation variable $W_{i_f n}$,

$$\begin{aligned} W_{i_f n} &= 1 \quad \text{if person } n \text{ knows the correct answer to item } i_f, \\ W_{i_f n} &= 0 \quad \text{if person } n \text{ does not know the correct answer to item } i_f. \end{aligned} \tag{17}$$

The conditional probability of $W_{i_f n}$ given response $U_{i_f n}$ is given by

$$\begin{aligned} &p(W_{i_f n} = 1|U_{i_f n} = 1, \theta_n, \boldsymbol{\xi}_{i_f}) \propto \Phi(a_{i_f}\theta_n - b_{i_f}), \\ &p(W_{i_f n} = 0|U_{i_f n} = 1, \theta_n, \boldsymbol{\xi}_{i_f}) \propto \gamma_{i_f}\big(1 - \Phi(a_{i_f}\theta_n - b_{i_f})\big), \\ &p(W_{i_f n} = 1|U_{i_f n} = 0, \theta_n, \boldsymbol{\xi}_{i_f}) = 0, \\ &p(W_{i_f n} = 0|U_{i_f n} = 0, \theta_n, \boldsymbol{\xi}_{i_f}) = 1. \end{aligned} \tag{18}$$

If a person gives an incorrect response to an item, (s)he is assumed not to know the answer; otherwise it would have been given.

Let $\boldsymbol{\delta}_{i_f} = (a_{i_f}, b_{i_f})$. A second augmentation variable, $\boldsymbol{Z}$, is defined conditionally on $\boldsymbol{W}$ as a truncated normally distributed variable (Albert, 1992):

$$Z_{i_f n} \mid W_{i_f n}, \theta_n, \boldsymbol{\delta}_{i_f} \sim \Phi(a_{i_f}\theta_n - b_{i_f}), \tag{19}$$

$$\begin{aligned} Z_{i_f n} &< 0 \quad \text{if } W_{i_f n} = 0, \\ Z_{i_f n} &\geq 0 \quad \text{if } W_{i_f n} = 1. \end{aligned} \tag{20}$$

In terms of the interpretation given above, if person n does not know the answer to item i_f the augmentation variable $Z_{i_f n}$ will be negative. Otherwise, it will be positive.

Because of these two types of data augmentation, the conditional posterior distributions of the ability and item discrimination and difficulty parameters can be obtained through a standard result from linear regression theory, which shows that the conditional posterior distributions of θ_n and $\boldsymbol{\delta}_{i_f}$ are (multivariate) normal,

$$\theta_n \mid \boldsymbol{z}_n, \boldsymbol{\delta} \sim N\left(\frac{\sum_{f=1}^{F} \sum_{i_f=1}^{I_f} a_{i_f}(z_{i_f n} + b_{i_f})}{1 + \sum_{f=1}^{F} \sum_{i_f=1}^{I_f} a_{i_f}^2}, \left(1 + \sum_{f=1}^{F} \sum_{i_f=1}^{I_f} a_{i_f}^2\right)^{-1} \right), \tag{21}$$

$$\boldsymbol{\delta}_{i_f} \mid \boldsymbol{z}_{i_f}, \boldsymbol{\theta}, \boldsymbol{\mu}_{\delta_f}, \boldsymbol{\Sigma}_{\delta_f} \sim \text{MVN}\big(\hat{\boldsymbol{\delta}}_{i_f}, \big(\boldsymbol{\Sigma}_{\delta_f}^{-1} + \boldsymbol{X}^T \boldsymbol{X}\big)^{-1}\big), \tag{22}$$

and

$$\hat{\boldsymbol{\delta}}_{i_f} = \big(\boldsymbol{\Sigma}_{\delta_f}^{-1} + \boldsymbol{X}^T \boldsymbol{X}\big)^{-1} \big(\boldsymbol{\mu}_{\delta_f} \boldsymbol{\Sigma}_{\delta_f}^{-1} + \boldsymbol{X}^T \boldsymbol{z}_{i_f}\big),$$

where $\boldsymbol{X} = (\boldsymbol{\theta}, -\mathbf{1}_N)$, and $\boldsymbol{\mu}_{\delta_f}$ and $\boldsymbol{\Sigma}_{\delta_f}$ are the elements of $\boldsymbol{\mu}_f$ and $\boldsymbol{\Sigma}_f$ corresponding to the discrimination and difficulty parameters. Note that $\hat{\boldsymbol{\delta}}_{i_f}$ is a precision weighted estimate composed of the family parameters $\boldsymbol{\mu}_{\delta_f}$ and $\boldsymbol{\Sigma}_{\delta_f}$ and an item-specific effect, which is the classical least-squares estimate. The family prior causes the item parameter estimates to 'shrink' towards their family means. The restriction of positive discrimination parameters can be applied by redrawing when the restriction is violated.

Guessing parameters γ_{i_f} are obtained in a separate step, conditional on $\boldsymbol{W}$, $\boldsymbol{U}$, and the draws of the second-level parameters. Let t_{i_f} be the number of persons who do not know the correct answer to item i_f and guess a response. The number of correct guesses, s_{i_f}, has a binomial distribution with parameters γ_{i_f} and t_{i_f}. We assume that $c_{i_f} = \text{logit}(\gamma_{i_f})$ has a normal prior with parameters $\boldsymbol{\mu}_{c_f|\delta_f}$ and $\boldsymbol{\Sigma}_{c_f|\delta_f}$. Combining a binomial likelihood with this normal prior results in a non-standard posterior distribution but importance sampling can be used to obtain draws from it (Gelman et al., 2004).

A.2. Drawing the Second-Level Parameters

The family parameters and effect parameters for the radicals are estimated using multivariate linear regression. The regression equation for item i_f is

$$\boldsymbol{\xi}_{i_f} = \boldsymbol{Q}_f \boldsymbol{\lambda} + \epsilon_{i_f} = \begin{bmatrix} 0 & .. & 1 & .. & 0 & \mathbf{0} & 0 & .. & 0 & .. & 0 \\ 0 & .. & 0 & .. & 0 & \boldsymbol{d}_f^T & 0 & .. & 0 & .. & 0 \\ 0 & .. & 0 & .. & 0 & \mathbf{0} & 0 & .. & 1 & .. & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\beta} \\ \boldsymbol{\mu}_c \end{bmatrix} + \epsilon_{i_f}, \tag{23}$$

where ϵ_{i_f} is our generic notation for a vector of normally distributed error terms with mean zero and covariance matrix $\boldsymbol{\Sigma}_f$. Parameter vector $\boldsymbol{\lambda}$ contains F family means for the discrimination parameters, R parameters for the effects of the radicals, and F family means for the guessing parameters. Note that the position of the 1s in indicator matrix $\boldsymbol{Q}_f$ depends on f. In the first row (= discrimination parameter) the position is f, whereas in the last row (= guessing parameter) the position is $F + R + f$.

In the representation in (23), each discrimination and guessing parameter is thus treated as an intercept (= family mean) plus an error term describing the variability around it. Likewise,

each difficulty parameter is treated as the sum of the effect parameters for the radicals in the family plus an error term.

Equation (23) is an example of what Griffiths and Valenzuela (2006) have called a set of "seemingly unrelated regressions". The term was coined by Zellner (1962) and refers to regression models in which each dependent variable is related to a possibly different set of explanatory variables. For correlated dependent variables, Zellner (1962) shows that the use of generalized least squares results in more efficient estimation of the regression coefficients than ordinary least squares.

Using the multivariate normal prior given in (7) the conditional posterior distribution of $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} \mid \boldsymbol{Q}, \boldsymbol{\Sigma}, \boldsymbol{\xi} \sim \text{MVN}\left(\hat{\boldsymbol{\lambda}}, \left(\boldsymbol{V}_0^{-1} + \sum_{f=1}^{F} I_f\left(\boldsymbol{Q}_f^T \boldsymbol{\Sigma}_f^{-1} \boldsymbol{Q}_f\right)\right)^{-1}\right), \tag{24}$$

where

$$\hat{\boldsymbol{\lambda}} = \left(\boldsymbol{V}_0^{-1} + \sum_{f=1}^{F} I_f\left(\boldsymbol{Q}_f^T \boldsymbol{\Sigma}_f^{-1} \boldsymbol{Q}_f\right)\right)^{-1} \left(\boldsymbol{V}_0^{-1}\boldsymbol{\lambda}_0 + \sum_{f=1}^{F} \sum_{i_f=1}^{I_f} \boldsymbol{Q}_f^T \boldsymbol{\Sigma}_f^{-1} \boldsymbol{\xi}_{i_f}\right),$$

is the generalized least squares estimator and $\boldsymbol{\Sigma}_f$ is a family covariance matrix drawn from its conditional posterior distribution, which is an inverse-Wishart described below. The draws of the family means of the difficulty parameters are constructed as the sum of the draws for their effect parameters; see (5).

When larger numbers of items per family are administered, the family-specific covariance matrices in the LICM-F can be estimated. Using the inverse-Wishart prior given in (8) the conditional posterior distribution of the matrices is equal to

$$\boldsymbol{\Sigma}_f | \boldsymbol{\xi}_f, \boldsymbol{\mu}_f \sim \text{inverse-Wishart}\left(\nu_0 + I_f, (\boldsymbol{S}_0 + \boldsymbol{S}_f)^{-1}\right), \tag{25}$$

with scale matrix $\boldsymbol{S}_f$ defined as

$$\boldsymbol{S}_f = \sum_{i_f=1}^{I_f} (\boldsymbol{\xi}_{i_f} - \boldsymbol{\mu}_f)(\boldsymbol{\xi}_{i_f} - \boldsymbol{\mu}_f)^T,$$

and $\boldsymbol{\mu}_f = \boldsymbol{Q}_f \boldsymbol{\lambda}$. However, estimation of family-specific covariance matrices may lead to estimates with large standard errors when the number of items per family is small. As an alternative, it then makes sense to assume equal covariance structures across families (LICM-C) and estimate their common covariance matrix as

$$\boldsymbol{\Sigma} | \boldsymbol{\xi}, \boldsymbol{\mu} \sim \text{inverse-Wishart}\left(\nu_0 + K, (\boldsymbol{S}_0 + \boldsymbol{S})^{-1}\right), \tag{26}$$

with

$$\boldsymbol{S} = \sum_{f=1}^{F} \sum_{i_f=1}^{I_f} (\boldsymbol{\xi}_{i_f} - \boldsymbol{\mu}_f)(\boldsymbol{\xi}_{i_f} - \boldsymbol{\mu}_f)^T.$$

The Gibbs sampling scheme iteratively draws from the distributions of $\boldsymbol{W}$, $\boldsymbol{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, $\boldsymbol{c}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\lambda}$.

References

Albert, J.H. (1992). Bayesian estimation of normal-ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*, 261–269.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–562.

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

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

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

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: a generalized linear and nonlinear approach*. New York: Springer.

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

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

Fox, J.-P. (2004). Multilevel IRT model assessment. In L.A. van der Ark, M.A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 227–252). London: Lawrence Erlbaum Associates.

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

Freund, P.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.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis*. New York: Chapman & Hall.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*, 241–251.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J.M. Bernardo, J. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 4: proceedings of the fourth Valencia international meeting* (pp. 169–193). Oxford: Oxford University Press.

Glas, C.A.W. (2010). Item parameter estimation and item fit analysis. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 269–288). New York: Springer.

Glas, C.A.W., & van der Linden, W.J. (2001). *Modeling variability in item parameters in item response models* (Research Report 01-11). Enschede, The Netherlands: Department of Educational Measurement and Data Analysis, University of Twente.

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

Glas, C.A.W., van der Linden, W.J., & Geerlings, H. (2010). Estimation of the parameters in an item-cloning model for adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 289–314). New York: Springer.

Griffiths, W.E., & Valenzuela, M.R. (2006). Gibbs samplers for a set of seemingly unrelated regressions. *Australian and New Zealand Journal of Statistics*, *48*, 335–351.

Heidelberger, P., & Welch, P.D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement items. *Journal of Educational Measurement*, *5*, 275–290.

Holling, H., Bertling, J.P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, *35*, 71–76.

Irvine, S.H., (2002). The foundations of item generation for mass testing. In S.H. Irvine & P.C. Kyllonen (Eds.) *Item generation for test development* (pp. 3–34). Mahwah: Lawrence Erlbaum Associates.

Janssen, R., Schepers, J., & Peres, 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.

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.

Johnson, V.E., & Albert, J.H. (1999). *Ordinal data modeling*. New York: Springer.

Laros, J.A., & Tellegen, P.J. (1991). *Construction and validation of the SON-R 5,5-17, the Snijders-Oomen non-verbal intelligence test*. Groningen: Wolters-Noordhoff.

Luecht, R.M. *Adaptive computer-based tasks under an assessment engineering paradigm*. Paper presented at the 2009 Graduate Management Admission Council Conference on Computerized Adaptive Testing, Minneapolis, Minnesota.

MacEachern, S.N., & Berliner, L.M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, *48*, 188–190.

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., & Levy, R. (2007). Bayesian psychometric modeling from an evidence-centered design perspective. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 839–865). Amsterdam: Elsevier.

Osburn, H.G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement*, *28*, 95–104.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*, 7–11. Available from http://CRAN.R-project.org/doc/Rnews/.

R Development Core Team (2009). R: A language and environment for statistical computing. Computer software manual. Vienna, Austria. Available from http://www.R-project.org.

Raftery, A.E., & Lewis, S. (1992). How many iterations in the Gibbs sampler? In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 4: proceedings of the fourth Valencia international meeting* (pp. 763–773). Oxford: Oxford University Press.

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

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

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

Roid, G., & Haladyna, T. (1982). *A technology for test-item writing*. New York: Academic Press.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

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.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*, 583–639.

Tanner, M.A. (1996). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*. New York: Springer.

Tellegen, P.J., & Laros, J.A. (1993). The construction and validation of a nonverbal test of intelligence: the revision of the Snijders-Oomen tests. *European Journal of Psychological Assessment*, *9*, 147–157.

van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386.

van der Linden, W.J., & Glas, C.A.W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*, 35–53.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association*, *57*, 348–368.

Manuscript Received: 23 MAR 2010
Final Version Received: 19 JUL 2010
Published Online Date: 17 MAR 2011