Computational Statistics (2000) 15:421–442

© Physica-Verlag 2000

Exploring the posterior of a hierarchical IRT model for item effects

Rianne Janssen and Paul De Boeck[1]
Department of Psychology, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium

Summary

A one-way ANOVA structure is imposed on the item difficulty and the item discrimination parameter of a two-parameter hierarchical IRT model for item effects. Bayesian estimation of the model is illustrated for the Metropolis-Hastings within Gibbs and the data augmented Gibbs procedure. The posterior of the hierarchical IRT model is explored with respect to the location of parameters and the uncertainty of these parameter estimates. The posterior correlations among parameters are shown to be due to trade-off effects among parameters either on the same parameter scales or on different parameter scales.

Keywords: Gibbs Sampler, Hierarchical Modeling, IRT, Item Effects, Posterior Correlations

[1]We thank the editor and two referees for helpful comments. The first author is a Postdoctoral Fellow of the Fund for Scientific Research - Flanders (Belgium). The construction of the test used in the example was funded by the OBPWO grant 93.05 of the Ministry of the Flemish Community, Department of Education, awarded to Paul De Boeck and Frans Daems.

1 Introduction

Item Response Theory (IRT) is a valuable tool for the analysis of the responses of a group of persons to a set of items. The core idea of IRT models is that the probability of a certain response can be described as a function of the person's position on a latent scale on the one hand, and one or more parameters characterizing the item on the other hand. Since the seminal work of Rasch (1960), numerous IRT models have been proposed (see e.g., van der Linden & Hambleton, 1997), including models for dichotomous as well as polytomous responses, and using either a unidimensional or a multidimensional latent space. Most of these models have found applications in psychological and educational measurement, where the IRT approach has proven to be useful in tackling problems such as test equating, item bias, and computer based assessment.

The present paper discusses an extension of a standard unidimensional IRT model for dichotomous responses, namely the two-parameter normal ogive model (Lord & Novick, 1968). If X_{vi} equals 1 if person v responds to item i correctly (or positively), and X_{vi} equals 0 otherwise, the probability of a correct response is then defined in this model as

$$P(X_{vi} = 1|\theta_v, \beta_i, \alpha_i) = p_{vi}(\theta_v, \beta_i, \alpha_i) = \Phi(\alpha_i(\theta_v - \beta_i)), \tag{1}$$

where $\Phi(.)$ denotes the probit transformation. θ_v refers to the person's position on the latent scale and is usually interpreted as an ability parameter. β_i is estimated on the same scale as θ_v and measures the item difficulty. Finally, α_i denotes the item discrimination parameter and represents the slope of the probit regression line relating the latent dimension to the success probability. It can be easily inferred from Equation (1) that the success probability increases, the more the person ability exceeds the item difficulty. However, the difference between these two parameters is weighted by the item discrimination parameter. For the two-parameter model to be identified, the origin and scale of the latent dimension have to be fixed.

Janssen, Tuerlinckx, Meulders, and De Boeck (in press) proposed a hierarchical extension of the two-parameter normal ogive model in which a one-way ANOVA structure is imposed on the item parameters. In this model, the items are partitioned on an apriori basis into K groups, each containing I_k items ($k = 1, , K$). Each item group k is characterized by a certain difficulty ξ_k on the latent scale and by a discrimination power ω_k. The difficulty parameters β_{ik} of the items within item group k are assumed to be normally distributed with mean ξ_k and variance σ_k^2:

$$\beta_{ik} \sim \mathrm{N}(\xi_k, \sigma_k^2). \tag{2}$$

Likewise, the item discrimination parameters α_{ik} of the items belonging to item group k are assumed to be normally distributed with mean ω_k and variance ν_k^2:

$$\alpha_{ik} \sim \mathrm{N}(\omega_k, \nu_k^2). \tag{3}$$

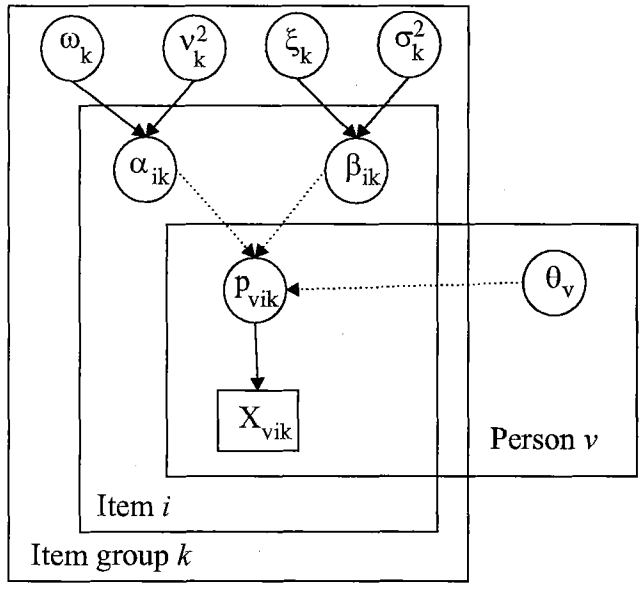

Figure 1: Graphical representation of the hierarchical IRT model for item effects.

Equations (2) and (3) imply that the item parameters β_{ik} and α_{ik} are considered to be random effects. At the item level, this may be interpreted as that the items of an item group are a random sample from some universe of items. However, for a given item, its difficulty and discrimination parameters are assumed to be sampled independently, i.e., there is no prior correlation assumed between the two item parameter distributions.

Combining Equations (1) to (3) leads to an IRT model with a hierarchical structure at the item side. The parameters ξ_k, ω_k, σ_k^2 and ν_k^2 are the so-called hyperparameters of the model. They imply a probabilistic specification of the item parameters α_{ik} and β_{ik}, with which the observable outcomes are modeled. A graphical outline of the model is presented in Figure 1 by means of a directed acyclic graph (see e.g., Spiegelhalter, Best, Gilks & Inskip, 1996). In such a graph, circles denote unobserved quantities and squares observed variables. The solid arrows refer to probabilistic relations like those implied by the distributional assumptions in the model, whereas the dotted arrows correspond to the deterministic, functional relations described in the model. The variables that are captured in the same polygon characterize the same real-world component.

Janssen et al. (1998) developed the hierarchical IRT model in the context of criterion-referenced measurement. In criterion-referenced measurement, the focus of interest is the performance of a person on a specific criterion, which is operationalized by looking at the performance on the sample of items

used to measure the criterion. In the hierarchical IRT model, the item groups refer to different criteria measured on the same latent scale. The hyperparameters characterize the criteria and are estimated on the basis of the items that are used to measure the criteria. The probability of success on a criterion can then be assessed by calculating the probability of success according to Equation (1), using the person ability parameter estimate (based on the performance on the items), but substituting the item parameters by the corresponding hyperparameters. This probability of success on the criterion can be considered to be an estimation of the mastery level of the person on the criterion. Using cutoff points on the probability scale, Janssen et al. (1998) were able to determine the proportion of students to be classified in different mastery categories on a set of educational attainment targets for reading comprehension.

Another application of the hierarchical IRT model may be found in the field of automated item generation (e.g., Embretson, 1998). When an item is automatically generated according to an item generation design, one is generally not able to assess with sufficient precision the value of the item parameters without empirical item calibration. However, using the hierarchical IRT model, one may be able to describe the distribution of the item parameters of items that are generated according to the same template. In some situations, the hyperparameters of this distribution may be used as a substitute for the corresponding individual item parameter estimates.

It is the purpose of the present article to further explore the properties of the hierarchical IRT model. First, it will be shown that the model can be estimated quite straightforwardly using a Bayesian estimation based on Markov chain Monte Carlo (MCMC) methods. In particular, two versions of the Gibbs sampler will be presented, namely Metropolis-Hastings within Gibbs (MHG) and data augmented Gibbs (DAG). Currently, there seems no agreement as to which method is more apt for IRT models. Patz and Junker (1999) strongly advocated MHG, whereas Albert (1992), Baker (1998), and Bradlow, Wainer and Wang (1998) used DAG. Second, the posterior of the hierarchical IRT model will be explored on the basis of the random sample of parameters from the posterior distribution obtained from the Gibbs sampler. Posterior inference will be focused on the location and uncertainty of the parameters and also on the covariation among the parameters. Throughout the paper, the same data set as in Janssen et al. (in press) will be used. This data set is introduced first.

2 Data set: A flemish test for reading comprehension

In 1996, the Parliament of the Flemish Community in Belgium issued a set of attainment targets specifying the expected basic competencies of children

Table 1: Description of the nine attainment targets.

k	Level of Processing	Text Type	I_k
1	retrieving	instructions	4
2		articles in magazines	6
3	structuring	study material	8
4		tasks in textbooks	5
5		comics	9
6		stories, novels	6
7		poems	8
8	evaluating	newspapers, textbooks, encyclopedia's	6
9		advertisements	5

who leave primary education. The Department of Education funded a project (see De Boeck, Daems, Meulders, & Rymenams, 1997) to devise a test for assessing the attainment targets of reading comprehension in Dutch (being the official language in Flanders). These attainment targets were conceptualized as the ability to process information in texts that are meaningful to children. Each attainment target was characterized by two factors: (1) the required level of processing, and (2) the type of text. Three levels of processing were distinguished, namely retrieving, structuring and evaluating information in a text. A variety of types of texts was specified. We focus on a test of 57 dichotomously scored items for the first nine attainment targets, which was administered to 539 children who were in the final year of primary education. Table 1 gives a description of the contents of the test and of the number of items that were used to measure each attainment target. It was assumed that the nine attainment targets refer to a common underlying reading ability dimension. However, it should be clear that this unidimensionality with respect to the latent continuum does not imply that all attainment targets are expected to be equal in difficulty.

As in the present application the number of items in each item group was relatively small (ranging from 4 to 9 items) to expect stable estimates of the within-group variances, the variance parameters of the distributions of β_{ik} and α_{ik} were kept constant over groups, resulting in a single σ^2 and a single ν^2 to be estimated.

3 Bayesian estimation using the Gibbs sampler

3.1 Priors and posterior

For a Bayesian estimation of the hierarchical IRT model, prior distributions are needed for the hyperparameters and the person ability parameter. For ξ_k and ω_k ($k = 1,..., K$) non-informative, flat priors were chosen. As the prior distribution for the variance parameters σ^2 and ν^2, an inverse chi-square distribution with one degree of freedom was chosen, because this distribution is mathematically convenient for variance parameters in normal models (e.g., Gelman, Carlin, Stern, & Rubin, 1995). For the person ability parameters θ_v a standard normal prior distribution was chosen. Fixing the mean and variance of the prior for the person ability distribution is necessary to make the hierarchical IRT model identifiable with respect to the location and scale of the latent dimension.

Given these prior distributions and the change in the original model with respect to the variance parameters, the joint posterior distribution becomes:

$$p(\alpha,\beta,\theta,\xi,\omega,\sigma^2,\nu^2|X) \propto p(\sigma^2)p(\nu^2)\left(\prod_{v=1}^{n} \mathrm{N}(\theta_v;0,1)\right)\times$$

$$\left(\prod_{k=1}^{K}\prod_{i=1}^{I_k} \mathrm{N}(\beta_{ik};\xi_k,\sigma^2)\right)\left(\prod_{k=1}^{K}\prod_{i=1}^{I_k} \mathrm{N}(\alpha_{ik};\omega_k,\nu^2)\right)\times \qquad (4)$$

$$\left(\prod_{v=1}^{n}\prod_{k=1}^{K}\prod_{i=1}^{I_k}(p_{vik}(\theta_v,\beta_{ik},\alpha_{ik}))^{x_{vik}}(1-p_{vik}(\theta_v,\beta_{ik},\alpha_{ik}))^{1-x_{vik}}\right).$$

In Equation (4), as well as in the following, leaving out a subscript for a parameter means that a vector of parameter values is considered with the vector running over the missing subscript. For example, β_k equals (β_{1k}, ..., β_{ik}, ..., $\beta_{I_k k}$) and β equals (β_1, ... , β_k, ..., β_K). The notation $N(Y;\mu,\tau^2)$ indicates that Y is normally distributed with mean μ and variance τ^2.

The posterior of the hierarchical IRT model consists of three parts: (a) the prior distributions, (b) the distributions with respect to the hierarchical structure imposed on the items, and (c) the likelihood of the data given the person and item parameters. Independence assumptions are made in all three parts: namely, firstly, independence of the priors, secondly, independence of the item parameter distributions, and, thirdly, independence of the observations within persons given θ_v (i.e., local independence) and of the observations across persons (i.e., experimental independence).

3.2 The Gibbs sampler

In order to obtain a sample from the joint posterior in Equation (4), we applied the Gibbs sampler, which is an iterative technique that belongs to the class of MCMC techniques (Gelman et al., 1995; Gilks, Richardson & Spiegelhalter, 1996; Tanner, 1996). Without going into details, the Gibbs sampler can be described as a modular approach to sampling from the posterior distribution, creating a Markov chain by successively sampling from a set of full conditional distributions, consisting of the distribution of a parameter or a set of parameters given the data and the previously drawn values for the other parameters. Once this Markov chain has converged to its equilibrium distribution, one samples from the posterior distribution.

A crucial step in the application of the Gibbs sampler is the derivation of the full conditional distributions. Happily, due to the various kinds of independence assumptions in the posterior, a lot of parameters cancel out when deriving the conditional distributions. For the hyperparameters this leads directly to conditional distributions in closed form. For ξ_k and ω_k, normal distributions are obtained with mean and variance equal to their maximum likelihood estimates:

$$\xi_k|\beta_k,\sigma^2 \sim \mathrm{N}\left(\frac{\sum_{i=1}^{I_k}\beta_{ik}}{I_k}, \frac{\sigma^2}{I_k}\right) \tag{5}$$

and

$$\omega_k|\alpha_k,\nu^2 \sim N\left(\frac{\sum_{i=1}^{I_k}\alpha_{ik}}{I_k}, \frac{\nu^2}{I_k}\right). \tag{6}$$

For σ^2 and ν^2, scaled inverse χ^2 distributions are obtained, with the scale parameter being function of the within-group differences between the item parameters and their corresponding hyperparameter:

$$\sigma^2|\beta,\xi \sim Inv-\chi^2\left(1+\sum_{k=1}^{K} I_k, \frac{\sum_{k=1}^{K}\sum_{i=1}^{I_k}(\beta_{ik}-\xi_k)^2+1}{1+\sum_{k=1}^{K} I_k}\right) \tag{7}$$

and

$$\omega^2|\alpha,\omega \sim Inv-\chi^2\left(1+\sum_{k=1}^{K} I_k, \frac{\sum_{k=1}^{K}\sum_{i=1}^{I_k}(\alpha_{ik}-\omega_k)^2+1}{1+\sum_{k=1}^{K} I_k}\right). \tag{8}$$

However, due to the probit transformation in Equation (1), the conditional distributions for the person and item parameters are not directly obtainable. For these parameters we will compare two versions of the Gibbs sampler.

3.3 Metropolis-Hastings within Gibbs

Chib and Greenberg (1995) describe the general form of the MHG, although they object to that name, given that the Gibbs algorithm can be seen as a special case of the Metropolis-Hastings algorithm. In MHG a single iteration of the Metropolis-Hastings algorithm is performed for the parameters whose conditional Gibbs distributions are not known in closed form. In this Metropolis-Hastings step, a candidate value for the to be sampled parameter is drawn from a proposal distribution. The candidate value is accepted with a probability that equals the minimum of 1 and the acceptance ratio. As will be shown below, the acceptance ratio is higher, when the probability of the candidate value under the posterior distribution (given the values of the other parameters) times the probability under the proposal distribution is higher than the corresponding product of probabilities for the previously sampled value. If the candidate value is not accepted, the previously sampled value is kept. The proposal distribution can be chosen to be symmetrical for the candidate value and the previously sampled value, so that the proposal distribution cancels out in the calculation of the acceptance ratio.

Following Patz and Junker (1999), we used symmetric, normal proposal distributions for θ_v and β_{ik} and a nonsymmetric, lognormal distribution for α_{ik}. The means of these distributions were set equal to the previously sampled value. We have chosen to sample β_{ik} and α_{ik} consecutively, while Patz and Junker (1999) sampled the two item parameters of each item together. Given these choices, the acceptance ratio at iteration t for a candidate value θ_v^* for the person ability parameter in comparison with the previously sampled value θ_v^{t-1} equals:

$$\frac{\left(\prod_{k=1}^{K}\prod_{i=1}^{I_k}\left(p_{vik}^*\right)^{x_{vik}}\left(1-p_{vik}^*\right)^{1-x_{vik}}\right) exp\left(-\frac{1}{2}\left(\theta_v^*\right)^2\right)}{\left(\prod_{k=1}^{K}\prod_{i=1}^{I_k}\left(p_{vik}^{t-1}\right)^{x_{vik}}\left(1-p_{vik}^{t-1}\right)^{1-x_{vik}}\right) exp\left(-\frac{1}{2}\left(\theta_v^{t-1}\right)^2\right)}, \tag{9}$$

where p_{vik}^* equals $p_{vik}(\theta_v^*, \beta_{ik}, \alpha_{ik})$ and p_{vik}^{t-1} is defined correspondingly. The accepancte ratio's for the item difficulty parameter β_{ik} and item discrimination parameter α_{ik} respectively equal:

$$\frac{\left(\prod_{v=1}^{N}\left(p_{vik}^*\right)^{x_{vik}}\left(1-p_{vik}^*\right)^{1-x_{vik}}\right) exp\left(-\frac{1}{2\sigma^2}\left(\beta_{ik}^*-\xi_k\right)^2\right)}{\left(\prod_{v=1}^{N}\left(p_{vik}^{t-1}\right)^{x_{vik}}\left(1-p_{vik}^{t-1}\right)^{1-x_{vik}}\right) exp\left(-\frac{1}{2\sigma^2}\left(\beta_{ik}^{t-1}-\xi_k\right)^2\right)} \tag{10}$$

and

$$\frac{\left(\prod_{v=1}^{N}\left(p_{vik}^{*}\right)^{x_{vik}}\left(1-p_{vik}^{*}\right)^{1-x_{vik}}\right) exp\left(-\frac{1}{2\nu^{2}}\left(\alpha_{ik}^{*}-\omega_{k}\right)^{2}\right)\frac{1}{\alpha_{ik}^{*}}}{\left(\prod_{v=1}^{N}\left(p_{vik}^{t-1}\right)^{x_{vik}}\left(1-p_{vik}^{t-1}\right)^{1-x_{vik}}\right) exp\left(-\frac{1}{2\nu^{2}}\left(\alpha_{ik}^{t-1}-\omega_{k}\right)^{2}\right)\frac{1}{\alpha_{ik}^{t-1}}}. \quad (11)$$

As described above, the structure of the acceptance ratio's is similar for the numerator and denominator. For θ_v and β_{ik} they describe the posterior probability, which consists of two parts: the likelihood of the model given the parameter estimates and the density under the (prior) distribution of the parameter. For α_{ik} a third part is added which is due to the lack of symmetry in the lognormal proposal distribution.

The variance parameters of the proposal distributions were set equal to 0.4, 0.08 and 0.04 for the θ_v, β_{ik} and α_{ik} distribution, respectively. These values were chosen on the basis of several preliminary runs of the Markov chain, so that average acceptance ratio's around 50% were obtained, which is recommended by Gelman, Roberts, and Gilks (1996) for univeriate parameter draws. For the final Markov chains (see below), the average acceptance ratio's were 56, 53, and 47% for the θ_v, β_{ik} and α_{ik} parameters, respectively. For reasons of numerical stability, the Metropolis-Hastings within Gibbs algorithm was programmed calculating the logarithm of the acceptance ratio's (as suggested by Patz & Junker, 1999).

3.4 Data augmented Gibbs

Albert (1992) showed that the conditional distributions for the Gibbs sampler for the standard two-parameter normal ogive IRT model can be derived analytically when the posterior distribution of the IRT model is augmented with latent data (Tanner, 1996). The data augmentation component in Albert's method consists in introducing a normally distributed latent response variable Z_{vik} with a mean that is a function of the person and item parameters:

$$Z_{vik} \sim \mathrm{N}\left(\alpha_{ik}\left(\theta_v-\beta_{ik}\right),1\right). \quad (12)$$

It is imposed that $Z_{vik} \leq 0$ when $x_{vik} = 0$, and that $Z_{vik} > 0$ when $x_{vik} = 1$. Hence, the probability that $Z_{vik} > 0$ equals the probability of a correct response given in Equation (1). Applying the data augmentation method to the hierarchical IRT model results in a truncated normal conditional distribution for the Z_{vik} as implied by Equation (12) and to normal distributions for the person and item parameters, namely:

$$\theta_v|Z_v,\alpha,\beta \sim \mathrm{N}\left(\frac{\sum_{k=1}^{K}\sum_{i=1}^{I_k}\alpha_{ik}\left(Z_{vik}+\alpha_{ik}\beta_{ik}\right)}{1+\sum_{k=1}^{K}\sum_{i=1}^{I_k}\alpha_{ik}^{2}}, \frac{1}{1+\sum_{k=1}^{K}\sum_{i=1}^{I_k}\alpha_{ik}^{2}}\right), \quad (13)$$

$$\beta_{ik}|\theta, Z_{ik}, \alpha_{ik}, \xi_k, \sigma^2 \sim \mathrm{N}\left(\frac{\alpha_{ik}\sum_{v=1}^{n}(\alpha_{ik}\theta_v - Z_{vik}) + \frac{\xi_k}{\sigma^2}}{n\alpha_{ik}^2 + \frac{1}{\sigma^2}}, \frac{1}{n\alpha_{ik}^2 + \frac{1}{\sigma^2}}\right), \quad (14)$$

and

$$\alpha_{ik}|\theta, Z_{ik}, \beta_{ik}, \omega_k, \nu^2 \sim \mathrm{N}\left(\frac{\sum_{v=1}^{n}((\theta_v - \beta_{ik})Z_{vik}) + \frac{\omega_k}{\nu^2}}{\sum_{v=1}^{n}(\theta_v - \beta_{ik})^2 + \frac{1}{\nu^2}}, \frac{1}{\sum_{v=1}^{n}(\theta_v - \beta_{ik})^2 + \frac{1}{\nu^2}}\right) \quad (15)$$

The means of each of these distributions can be written as a precision-weighted average of a latent residual and of a location parameter given by the prior mean (for θ_v) or by the corresponding hyperparameter (for α_{ik} and β_{ik}). The latent residual for each parameter can be derived by rewriting Equation (12) as $Z_{vik} = \alpha_{ik}(\theta_v - \beta_{ik}) + \varepsilon_{ik}$ with $N(\varepsilon_{ik}; 0, 1)$ and by filling in the fixed parts of the equation given the previously sampled values of the other parameters. The variance parameter of each normal distribution equals the inverse sum of the precisions of the latent residual and of the prior (for θ_v) or of the hierarchical distribution (for α_{ik} and β_{ik}).

3.5 Chains and convergence for the two Gibbs samplers

Five Markov chains of 2000 iterations were run for both versions of the Gibbs sampler. The first 1000 iterations were used for burn-in. The sampling order of the conditional distributions within one iteration was Z_{vik} (for DAG only), θ_v, β_{ik}, α_{ik}, ξ_k, ω_k, σ^2, and ν^2. For each chain, the starting values for the θ_v and the β_{ik} were drawn from the standard normal distribution and for α_{ik} from the uniform distribution with a minimum of 0 and a maximum of 2. The starting points for the hyperparameters consisted in their maximum likelihood estimates based on the starting values for the item parameters. The starting values for the DAG and MHG algorithm were the same.

Convergence of the Markov chain was checked for each model parameter on the basis of the measure $\sqrt{\widehat{R}}$ proposed by Gelman and Rubin (1992). This convergence measure is based on the idea that chains, running from different starting points in the parameter space, have reached convergence at the point where the variance of the sampled parameter values between the chains approximates the variance of the sampled parameter values within the chains. Formally, $\widehat{R}$ quantifies the estimated ratio of between-chain variation to within-chain variation of a single parameter. $\sqrt{\widehat{R}}$ has to be smaller than 1.1 (or at least 1.2, see Kass, Carlin, Gelman, & Neal, 1997) for all parameters of the model if one requires the MCMC algorithm to be converged. Our model

included 673 parameters, being 539 θ_v, 57 α_{ik} and β_{ik}, 9 ξ_k and ω_k, and finally σ^2 and ν^2.

When $\sqrt{\widehat{R}}$ was calculated for the last 1000 iterations of the five chains, there were no real differences in convergence between DAG and MHG. For the DAG algorithm, only 5 parameters showed $\sqrt{\widehat{R}}$ values being larger than 1.1 (but all smaller than 1.2), being the α_{ik} and β_{ik} parameters of two items and one α_{ik} of another item. For the MHG algorithm, 7 parameters showed nonconvergence according to the 1.1 criterion, being 5 β_{ik} and 2 α_{ik} parameters. However, for one item $\sqrt{\widehat{R}}$ was 1.35 for β_{ik} and 1.29 for α_{ik}. It turned out that the α_{ik} for this item drifted towards zero for each of the five chains, causing the β_{ik} to wander around in a large area of the latent scale. This item had a proportion of correct answers of .52, which is around the .50 probability that a two-parameter model with a zero discrimination predicts. More importantly, this nonconvergent item had the lowest discrimination according to DAG. In fact, all the sampled α_{ik} values with MHG were somewhat smaller than the sampled values with DAG, probably reflecting the fact that the lognormal distributions advantages small values. The smaller values of the α_{ik} in the MHG also resulted in a broader range for the sampled values of β_{ik} on the latent scale in comparison with DAG. It will be shown that this trade-off between α_{ik} and β_{ik} can also be found for the drawn parameter values of α_{ik} and β_{ik} of one item during Gibbs sampling, also with DAG.

When the lognormal proposal distribution for the α_{ik} was replaced with a normal distribution truncated at zero to avoid negative α_{ik} in another MHG implementation, no drift of an α_{ik} towards zero was observed. In fact, in comparison with the DAG procedure, the sampled values of α_{ik} of the latter MHG implementation were somewhat larger, this time resulting in a shrinkage of the difficulty scale in comparison with the DAG results. Note that Patz and Junker (1999) implemented the lognormal distribution for α_{ik} in a MHG were the α_{ik} and β_{ik} of one item were sampled together. This may prevent (low) α_{ik} to drift to zero.

4 Exploring the posterior

The exploration of the posterior is described in two parts. First, posterior inferences with respect to the location and uncertainty of the parameters are discussed. In order to make inferences about the location of the parameters of the hierarchical IRT model, we calculated expected aposteriori estimates by taking the corresponding sample averages from the draws from the posterior provided by the Gibbs sampler. The posterior standard deviation of the sampled values gives a measure of the posterior uncertainty of a parameter, which indicates to what degree the data are informative with respect to the quantity being estimated. Both the posterior mean and the posterior standard deviation can be considered explorations of the marginal posterior of a

Table 2: Summary statistics of the posterior means and posterior standards deviations of the item and person parameters.

	Mean			SD		
	Min	Max	Mean	Min	Max	Mean
β_{ik}	-5.18	4.64	-.96	.09	1.08	.32
α_{ik}	.16	.70	.46	.04	.11	.07
θ_v	-2.60	2.24	-.01	.34	.48	.37

parameter. In the second part, posterior correlations among parameters are investigated and explained. Exploring these posterior correlations is a step towards investigating the joint posterior of the parameters. They show in what degree the estimation of different parameters is interdependent.

The exploration of the posterior is done for two reasons. First, the exploration may render further insight in the hierarchical IRT model for item effects. Although the model is explored on the basis of only one data set, general properties of the posterior of the model can be inferred when the obtained results can be explained on the basis of model equations and Gibbs conditionals. Second, the exploration of the posterior aims at further investigating the data set of the Flemish test for reading comprehension. Janssen et al. (in press) used the posterior estimates of the hierarchical IRT for the test on reading comprehension to make inferences on the degree in which the nine attainment targets were mastered by a sample of students. In the present paper, some other aspects, like the differences between the attainment targets, will be looked at.

All the results below are based on the final 1000 draws from the posterior for the five chains of the DAG algorithm. The results of the MHG algorithm were comparable, and are therefore not further discussed.

4.1 Location and uncertainty

4.1.1 Item and person parameters

Table 2 gives some summary statistics for the posterior means and standard deviations of the item and person parameters. The item difficulty parameters covered quite a large range on the latent scale, with a mean location of -.96. Given that the origin of the scale was fixed at the mean of the prior for θ_v, the latter finding implies that on the average the items were relatively easy for the respondents. The posterior uncertainty about the location of the β_{ik} varied substantially. In fact, the posterior standard deviation of the β_{ik} showed a curvilinear relationship with the posterior mean, with item parameters in the middle range of the scale having the lowest posterior variance. This finding

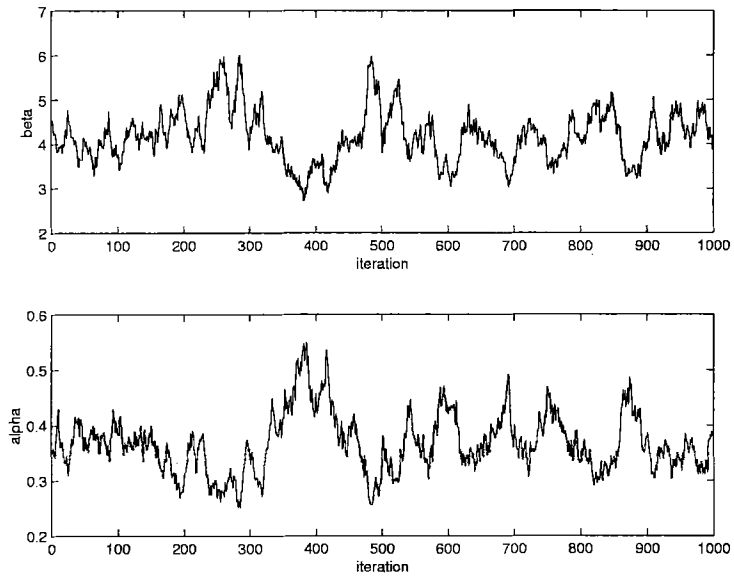

Figure 2: Trace plot of chain 1 (after burn-in) for β_{14} and α_{14}.

is in line with the maximum likelihood property for standard IRT models where the standard error of estimation is the largest for extreme difficulty parameters. As a visual illustration of the posterior uncertainty, the top of Figure 2 gives a trace plot of the 1000 final draws of one chain for β_{14} (item 1 of attainment target 4) with a posterior mean of 4.65. This chain shows a wavy pattern, covering a large range of the latent scale. This high level of posterior uncertainty for the extreme β_{ik} can be explained on the basis of the conditional distributions of the Gibbs sampler. For the MHG method, the likelihood of items with more extreme difficulty parameters is not informative as to the precise location of the items. By consequence, the parameter draws may wander around in a relatively large subrange of the latent scale. For the DAG method, the latent data Z_{vik} are coming for all persons from a larger range of the truncated normal distributions (being the positive end of the latent scale for the easy items and the negative end for the difficult items) making the occurrence of extreme Z_{vik} values more likely, which leads to more variability in the location of the conditional distribution of β_{ik} (and α_{ik}). Moreover, having once arrived in a certain region for the β_{ik} parameter, the interdependence between β_{ik}, Z_{vik}, θ_v, and α_{ik} (as shown below) does not make it easy to leave that region. For example, sampling an extreme β_{ik} also makes the probability of sampling extreme Z_{vik} more likely.

The posterior means and standard deviations of the item discrimination parameters α_{ik} varied in a less extent than those of the β_{ik}. The posterior means of the α_{ik} did not correlate over items with the posterior means of the

β_{ik}, although the larger values of α_{ik} were all observed in the middle range of the difficulty scale. For items with extreme difficulty, the trace plots of the sampled values of the α_{ik} were also quite wavy, as is illustrated in the bottom part of Figure 2 for the same item and chain as in the top part.

The posterior means of the person ability parameter θ_v varied in a smaller range of the latent scale than the posterior means of the β_{ik}. The posterior standard deviations of the θ_v showed a curvilinear relationship with the posterior mean, reflecting the uncertainty in measurement at the ends of the scale. Like for the β_{ik}, this finding can be derived from the conditional distributions of the Gibbs sampler.

4.1.2 Hyperparameters

Table 3 shows the postererior means and standard deviations of the hyperparameters ξ_k and ω_k. The variation among the location of the ξ_k was not very high. In order to investigate the differences between the item groups for ξ_k, paired comparisons among the ξ_k were calculated for each draw from the posterior. The the 95% posterior intervals of only three pairs of ξ_k did not include zero, namely ξ_9 - ξ_8, ξ_9 - ξ_4, and ξ_5 - ξ_8. These 'significant' comparisons refer to the difference between the easiest and the two most difficult attainment targets and to the difference between the second easiest with the most difficult attainment target. The posterior standard deviations of the ξ_k ranged from .63 to .88, and, hence, are generally larger than the posterior standard deviations for the β_{ik}. This is a normal observation for multilevel models: the higher in the hierarchy, the more posterior uncertainty. The posterior standard deviations of the ξ_k correlated -.97 with the number of items measuring the attainment target. In fact, it can be derived from Equation (5) that the posterior uncertainty for ξ_k is function of the posterior uncertainty of the mean of the β_{ik} and of the variance of this mean, being equal to σ^2 divided by the number of items.

The posterior mean of the discrimination powers ω_k of the attainment targets (see Table 3) varied between .42 and .58 for ω_1 to ω_8. Attainment target 9 showed a somewhat lower posterior mean of .32. All 95% posterior intervals for the paired comparisons among the ω_k included zero. The posterior standard deviations of the ω_k were almost linearly (r = -.96) related to the number of items measuring the attainment target, as was the case for the ξ_k.

The posterior mean of the variance σ^2 of the item difficulties within the attainment targets was 3.06 with a posterior standard deviation of .72. This relatively high within-group variance indicates that there is a high overlap in the item difficulty distributions for the nine item groups (as also the paired comparisons among the ξ_k indicated). The posterior mean for the within-group variance ν^2 of the α_{ik} was .05 with a posterior standard deviation of .01. This result indicates that the within-group variation among the items in α_{ik} was not so large.

Table 3: Posterior means and standard deviations of ξ_k and ω_k of the attainment targets.

	ξ_k		ω_k	
k	M	SD	M	SD
1	-.90	.88	.58	.12
2	-1.12	.72	.54	.10
3	-1.24	.63	.45	.08
4	.06	.81	.44	.10
5	-1.44	.60	.42	.08
6	-.70	.74	.47	.09
7	-1.07	.63	.47	.08
8	.57	.73	.48	.09
9	-2.53	.86	.32	.10

4.2 Posterior correlations

In maximum likelihood estimation, the estimation of the covariation among parameter estimates consists in the tedious task of calculating and inverting Hessian matrices. In the Bayesian MCMC approach, the covariation among parameter estimates can be estimated straightforwardly by calculating the correlation between draws of model parameters for each chain. The implementation structure of the Gibbs sampler in our example allows to estimate the posterior correlation among parameters for each of the five chains. In the following, the mean of these five correlations is used as the estimator of the posterior correlation. Another advantage of the MCMC approach is that the posterior correlations can be related to the conditional distributions of the Gibbs sampler, enhancing insight in these correlations, and, hence, in the model itself. In the Gibbs sampling procedure, each draw from a conditional distribution is essentially a random process. However, the parameters of these conditional distributions are function of a specific subset of previously sampled model parameters. This dependency can cause posterior correlations among parameters. Parameters that play a role in the location and/or scale of the conditional distribution of a particular parameter are those parameters on which the parameter is directly dependent, or with which the particular parameter directly shares a node in the graphical representation of the hierarchical IRT model in Figure 1. For example, the conditional distribution of the hyperparameter ξ_k is only determined by the β_{ik} of its item group and by σ^2. The conditional distribution of the β_{ik} is function of the hyperparameters ξ_k and σ^2, but also of the parameters θ_v and α_{ik} and of the (latent) data involved in the success probability of the item.

The posterior correlations of the hierarchical IRT model are due to two

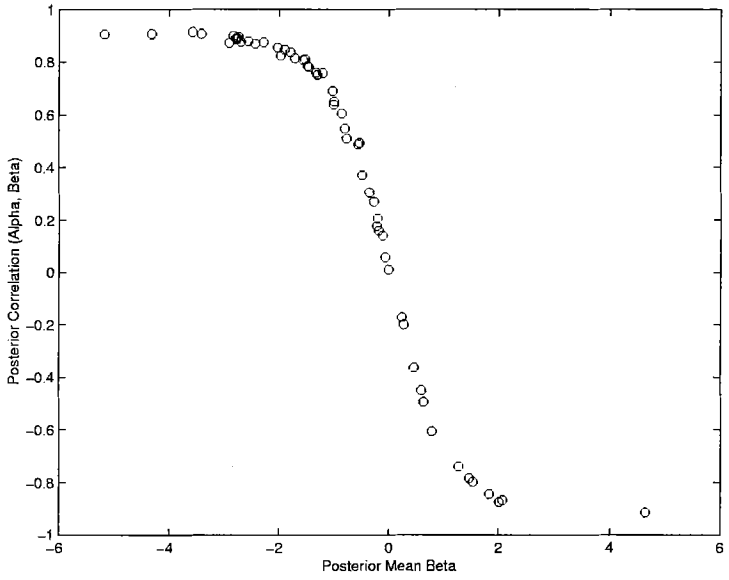

Figure 3: Plot of the posterior mean of β_{ik} against the posterior correlation between β_{ik} and α_{ik}.

kinds of trade-off effects among parameters. Trade-off effects can occur between two sets of parameters measured on a different scale, like between β_{ik} and α_{ik}. Trade-off effects can also refer to the covariation of draws of parameters on the same parameter scale, like between β_{ik} and ξ_k, or between α_{ik} and ω_k. These two kinds of trade-off are now discussed in turn.

4.2.1 Trade-offs between different parameter scales

The posterior correlation between β_{ik} and α_{ik} varied between -.91 and .92. As shown in Figure 3, this posterior correlation covaried strongly with the posterior mean of β_{ik}. The posterior correlation between β_{ik} and α_{ik} was positive for easy items and negative for difficult items. The more extreme the item difficulty, the higher the absolute value of the correlation, with zero correlations in the middle of the latent scale. There was no meaningful covariation between the posterior correlation between β_{ik} and α_{ik} and the posterior mean of α_{ik}. A visual illustration of a high posterior correlation is presented in Figure 2. The wavy patterns for β_{ik} are mirrored by the reversed wavy patterns for α_{ik}. For this chain, the posterior correlation between β_{ik} and α_{ik} was -.92. Note that the size of this posterior correlation is influenced by the autocorrelations among successive draws of β_{ik} and among successive draws of α_{ik}. When the chains of β_{ik} and of α_{ik} are both corrected for autocorrelation of lag(1), the posterior correlation between the two chains

drops to -.48.

The posterior correlations between β_{ik} and α_{ik} are the Bayesian counterpart of the covariances found in ML estimates of item parameters of IRT models (Wingersky & Lord, 1984). The covariation in the estimation of the difficulty and the discrimination parameter of IRT models was already observed by Lord (1975) who called it " a distinctly unpleasant property" (p. 206). Stocking (1989) explained that the correlation between errors of sample estimates of item discrimination and item difficulty is not a consequence of a particular estimation procedure, but of the fact that the IRT model has a constrained form so that " a particular misestimation of one item parameter causes an automatic and predictable adjustment in the estimates of the other parameters in order to fit the available data as well as possible" (p. 9). Hence, the correlation between estimates of item difficulty and item discrimination is due to a trade-off effect.

Take for example an item with real item difficulty equal to +3. Such a value implies that all respondents will have low success probabilities on this item. An overestimated difficulty of +4, resulting in even lower success probabilities as the differences between θ_v - β_{ik} get smaller, can be compensated by a lower value of the discrimination parameter, as α_{ik} premultiplies the difference between θ_v - β_{ik} in model for the success probability (see Equation 1). Hence, for difficult items (with a positive β_{ik}), one can expect a negative correlation between estimates of difficulty and of item discrimination. In the same vein, one can explain why the item difficulty and item discrimination of easy items (with a negative β_{ik}) are positively correlated.

The trade-off between β_{ik} and α_{ik} can also be found in the Gibbs conditionals. For MHG, the under- and overestimation of β_{ik} will cause the likelihood to be greater in a particular direction for possible values of α_{ik}. Likewise, the under- or overestimation of α_{ik} at iteration t, may increase the likelihood for specific subranges for the β_{ik} parameter at iteration$t+1$. The transfer effect from α_{ik} to β_{ik} may however be less strong, because when β_{ik} is sampled at iteration $t+1$, also the values of the person ability parameters and hyperparameters have changed. For DAG, the effect of sampled values of β_{ik} on the location of the conditional distribution of α_{ik} has to be investigated in connection with the values of the latent variables Z_{vik}, as can be seen in Equation (15). For easy items, the Z_{vik} will generally be positive for the respondents, while for difficult items the Z_{vik} will generally be negative. The expected value of Z_{vik} for each person will be a linear function of θ_v. Suppose that the sampled value of β_{ik} of a difficult item gets smaller (more negative), then the difference between θ_v and β_{ik} gets smaller too, but as the difference is generally multiplied by a negative Z_{vik} (because of the item being difficult), the location of the α_{ik} distribution will generally increase (cf. Equation 15), leading to the negative correlation between draws of α_{ik} and of β_{ik}. As for MHG, the way back from a drawn value of α_{ik} at iteration t to the location of the conditional distribution of β_{ik} at iteration $t+1$ is less

Table 4: Range of the posterior correlations due to trade-off on the same parameter scale.

Scale	Parameters	Min	Max
Difficulty	β_{ik}, ξ_k	.03	.30
	β_{ik}, σ^2	-.25	.22
	ξ_k, σ^2	-.12	.02
	Mean θ_v, Mean ξ_k	—	.19
Discrimination	α_{ik}, ω_k	.10	.25
	α_{ik}, ν^2	-.07	.17
	ω_k, ν^2	.01	.05

influential, because of the intermediate sampling of other (hyper)parameters that determine the conditional distribution of β_{ik} as well. So the correlation between α_{ik}^t and β_{ik}^{t+1} will generally be smaller than the correlation between α_{ik}^t and β_{ik}^t .

The trade-off between β_{ik} and α_{ik} may also cause a (smaller) trade-off between the hyperparameters ξ_k and ω_k of the same item group. In fact, it was found that the posterior correlation between ξ_k and ω_k varied between -.01 and .08. Despite these low values, the mean posterior correlation correlated -.92 with the posterior mean of ξ_k: lower posterior means of ξ_k corresponded with higher mean correlations. The mean posterior correlation between ξ_k and ω_k did not correlate with the posterior mean of ω_k. Hence, like for the trade-off between α_{ik} and β_{ik}, the small trade-off between the hyperparameters ξ_k and ω_k is a function of difficulty.

4.2.2 Trade-offs on the same parameter scale

The posterior correlations between parameters measured on the same parameter scale indicate that the uncertainty in the location of one parameter set covaries with the uncertainty in the location of another parameter set. Moreover, the uncertainty in the location of one parameter set can give rise to fluctuations in the estimates of the variance parameter. These trade-offs in location and in variance can be found both on the difficulty and on the discrimination scale. Table 4 gives the range of the posterior correlations due to trade-off on the same parameter scale.

Trade-Offs on the Difficulty Scale. The posterior correlations between β_{ik} and ξ_k were all positive. These positive posterior correlations are due to the fact that the two parameters determine the location of each other's conditional distribution. The uncertainty in the location of the β_{ik} therefore covaried with the uncertainty in the location of the ξ_k. The strength of the posterior correlations covaried in a meaningful way with the extremeness of

the parameters. The posterior correlation between β_{ik} and ξ_k was stronger for extreme β_{ik}. Hence, the mean of the conditional distribution of ξ_k is influenced more heavily by the translations on the scale by extreme items in the item group. The effect of ξ_k on the location parameter of the conditional distribution of β_{ik} depends on σ^2. When σ^2 is large, the effect of ξ_k on the mean location of β_{ik} diminishes.

As a second example of the covariation in the uncertainty in the location of two parameter sets on the difficulty scale, the posterior correlation between θ_v and ξ_k is investigated. As the θ_v and the ξ_k are measured on the same latent scale, it can be expected that when the person ability parameters become on average somewhat larger, the mean item difficulty (and, hence, the mean of ξ_k) increases. Accordingly, it was found that the mean of θ_v over all 539 subjects for each draw showed a mean correlation of .19 with the mean ξ_k over the nine item groups for each draw.

The uncertainty in the location of the β_{ik} and of the ξ_k were both associated with the variance parameter σ^2 (see Table 4). This covariation was stronger for the more extreme parameter values of β_{ik} and of ξ_k. The posterior correlation between β_{ik} and σ^2 correlated .92 with the posterior mean of β_{ik}: fluctuations in the extreme negative values of β_{ik} correlated negatively with draws of σ^2, whereas fluctuations in extreme positive values of β_{ik} correlated positively with draws of σ^2. The same was found for the posterior correlation between ξ_k and σ^2 which correlated .86 with the posterior mean of ξ_k. The influence of β_{ik} and by ξ_k on σ^2 can also be inferred from the Gibbs conditional for the variance parameter.

Trade-Offs on the Discrimination Scale. For the discrimination scale, the uncertainty in the location of the α_{ik} and of the ω_k were positively correlated (see Table 4). Again, this trade-off effect was stronger for the extreme parameters. The posterior correlation between α_{ik} and ω_k correlated .57 with the posterior mean of α_{ik}. Hence, the location of ω_k was more strongly influenced by the larger α_{ik}. On the other hand, the posterior correlation between α_{ik} and ω_k averaged over the items belonging to the same item group correlated .60 with the posterior mean of ω_k. Hence, the location of α_{ik} was more strongly influenced by the larger ω_k.

The uncertainty in the location of α_{ik} and ω_k also covaried with the variance parameter on the discrimination scale. The posterior correlation between between ν^2 and α_{ik} correlated .83 with the posterior mean of α_{ik}. The posterior correlation between ν^2 and ω_k were low and did not covary with the posterior mean of ω_k. The low posterior correlations are probably due to the fact that the ω_k did not differ much among the attainment targets.

5 Discussion

The present paper illustrated two advantages of a Bayesian approach. First, estimation of models becomes relatively easy, even for complex models, like

the hierarchical IRT model for item effects in which a one-way ANOVA structure is imposed on the item parameters. Both the DAG and MHG versions of the Gibbs sampler worked well for this model. Second, measures of uncertainty of parameter estimates and of covariation among parameter estimates are generated almost automatically. This is quite an advantage in comparison with the calculation of a variance-covariance matrix for maximum likelihood estimators. Moreover, the results of the (co)variance of parameter estimates can be linked to the structure of the Gibbs conditionals, making the estimation of the model more insightful. The posterior correlations indicate that the uncertainty of the parameter estimates is not a pure random process, but that the uncertainty in one parameter covaries with the estimation and uncertainty in other parameters.

The exploration of het posterior of the hierarchical IRT model for item effects with respect to the location and uncertainty of the parameters first showed that the between-group differences among the items in difficulty were small in comparison with the within-group differences in the Flemish test for reading comprehension. This was evidenced by the paired comparisons among the ξ_k and by the large estimate of σ^2. With respect to discrimination, the item groups were even more homogeneous. These results imply that the classification of reading comprehension items along the scheme of the attainment targets does not explain a large proportion of the variation among the items. Other factors may succeed better in explaining item difficulty. Second, it was shown, as could be expected, that the reliability of the estimation of the difficulty and discrimination of the attainment targets is a function of the number of items in the item group. Finally, the posterior correlations between β_{ik} and ξ_k and between α_{ik} and ω_k showed that the estimation of the location of the hyperparameters covaried more strongly with the estimation of the more extreme item parameters in the item group. Hence, the reliability of the hyperparameters ξ_k and ω_k is also function of the homogeneity (in the sense of no extreme items) of the items in the item group.

The exploration of the posterior correlations were helpful in explaining the uncertainty in the location of paramaters. It was learned that there is a tradeoff between the uncertainty in the location of parameters on the difficulty scale and the uncertainty in the location of parameters on the discrimination scale. This was evidenced by the posterior correlations between β_{ik} and α_{ik} and in a less extent between ξ_k and ω_k. Note that the posterior correlation among the β_{ik} and α_{ik} parameter estimates for one item is totally different from the correlation over items between the posterior means of β_{ik} and α_{ik}. It was shown that the posterior means of β_{ik} and α_{ik} were uncorrelated over items, indicating that the strenght of the relationship between an item and the latent scale is not related with the position of the item on the latent scale.

The covariation in the uncertainty of the location of parameters on the same parameter scale (and their implications for the estimated within-group

variance) are a consequence of the model with an explicit dependence between the item parameters and their corresponding hyperparameters. The uncertainty in the location of the θ_v and of the ξ_k will influence the uncertainty with which the degree of mastery is measured in the hierarchical IRT model as this degree of mastery is determined on the basis of the difference between θ_v and ξ_k. However, the uncertainty in mastery may be less affected by the uncertainty in the locations of those parameters. This might be concluded from the marginal posteriors, given that on the average the translations of the ξ_k on the latent scale covaried with translations of the θ_v on the latent scale.

As a conclusion, we think that the exploration of the posterior beyond the marginal posteriors for each parameter set can be helpful both for model understanding and for explaining the uncertainty in the location of the parameters.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251-269

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169.

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

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician, 49*, 327-335.

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

Embretson, S. E. (1998). A cognitive design system approach to generation valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

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

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting* (pp. 599-608). New York: Oxford.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science, 7*, 457-511.

Gilks, W., Richardson, S. & Spiegelhalter, D. (eds.) (1996). *Markov Chain Monte Carlo in practice.* New York: Chapman & Hall.

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

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. N. (1997). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician, 52,* 93-100.

Lord, F. M. (1975). The 'ability' scale in item characteristic curve theory. *Psychometrika, 40,* 205-217.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., & Inskip, H. (1996). Hepatitis B: a case study in MCMC methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.). *Markov chain Monte Carlo in practice* (pp. 21-43). New York: Chapman & Hall.

Stocking, M. L. (1989). *Empirical estimation errors in item response theory as a function of test properties* (Research Report 89-5). Educational Testing Service, Princeton, NJ.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions (3rd ed.).* New York: Springer.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer.

Wingersky, M. S. & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Methods, 8,* 347-364.