Statistica Neerlandica (2005) Vol. 59, nr. 1, pp. 95–106

Bayesian modification indices for IRT models

Jean-Paul Fox* and Cees A. W. Glas†

Department of research methodology, measurement and data analysis, Twente University, 7500 AE Enschede, The Netherlands

Bayesian modification indices are presented that provide information for the process of model evaluation and model modification. These indices can be used to investigate the improvement in a model if fixed parameters are re-specified as free parameters. The indices can be seen as a Bayesian analogue to the modification indices commonly used in a frequentist framework. The aim is to provide diagnostic information for multi-parameter models where the number of possible model violations and the related number of alternative models is too large to render estimation of each alternative practical. As an example, the method is applied to an item response theory (IRT) model, that is, to the two-parameter model. The method is used to investigate differential item functioning and violations of the assumption of local independence.

Key Words and Phrases: Bayesian modification indices, Markov chain Monte Carlo, item response models.

1 Introduction

Model selection problems often concern choosing the most appropriate model from among a set of possible choices. However, a selection problem also occurs when just one specific model M_0 has been proposed, and the choice is between accepting or rejecting model M_0. In many instances, this model rejection problem (see, BERNARDO and SMITH (1994)) can be phrased as a hypothesis-testing problem, for example, the test of the simple hypothesis that a certain model parameter equals zero. In that case, the model rejection problem corresponds to testing certain assumptions of a specific model, say the null-model.

Consider a parametric framework $\{p(\lambda \mid \mathbf{y}), \lambda \in \Lambda\}$, where $p(\lambda \mid \mathbf{y})$ is the posterior distribution of λ given the data $\mathbf{y}$. Further, model M_0 corresponds to $p(\mathbf{y} \mid \lambda_0)$. The focus is on a point null hypothesis of the form $\lambda = \lambda_0$ and an alternative hypothesis $\lambda \neq \lambda_0$. It will be assumed that there is no reason to favour $\lambda = \lambda_0$ above $\lambda = \lambda_1$ for any value of λ_1 in the neighborhood of λ_0. To perform a

*G.J.A.Fox@edte.utwente.nl
†C.A.W.Glas@edte.utwente.nl

© VVS, 2005. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

Bayesian test of significance of the null hypothesis $\lambda = \lambda_0$ at a level of significance α, a credible interval is constructed from the posterior distribution and the null hypothesis is rejected if and only if λ_0 is outside this interval. This approach was introduced by LINDLEY (1965, section 5.6) and can also be found in ZELLNER (1971, section 10.2). The approach is based on the idea that the posterior distribution for a parameter λ is a basis for expressing beliefs about possible values of λ. If the value $\lambda = \lambda_0$ is in a region in which the posterior probability density is not high, this leads to the suggestion that this value for λ is not credible.

The primary purpose of modification indices is not so much testing the model, but providing diagnostic information on model fit for multi-parameter models. In a frequentist framework, in principle, model fit can be investigated by defining more general alternative models and using the likelihood ratio (LR) test to evaluate the seriousness of the violation. In multi-parameter models, however, the number of model violations can become very large. For example, item response theory (IRT) models describe the responses of students to test items, and model fit can be violated for every item and every person in many different ways. It is not practical to estimate an alternative model and compute an LR statistic for every one of these violations. Therefore, the initial fit analysis in IRT is often based on the analysis of residuals. One of the problems of the analysis of residuals, however, is that the presence of too large residuals does not automatically lead to the identification of the source of the misfit. An often used alternative is based on the Lagrange Multiplier (LM) test (AITCHISON and SILVEY, 1958), and the equivalent efficient score test (RAO, 1947). In the applications in the field of IRT (see GLAS (1998, 1999)) the null model is the IRT model and the alternative model implies model violations such as differential item functioning, violation of the assumed item characteristic functions and violation of local stochastic independence. The LM tests are computed using the parameter estimates under the null model. Therefore, a plethoria of model violations for all items can be assessed as a by-product of one single estimation run.

In the present article, this approach will be generalized to a Bayesian framework. It will be shown that an MCMC algorithm for estimating the parameters of the null-model can be extended to estimate the marginal posterior distribution of the added parameter λ, denoted as the Bayesian Modification (BM) distribution. Obviously, the marginal posterior distribution of λ is unknown but a credible interval is needed to perform the significance test of the hypothesis $\lambda = \lambda_0$. The extra step in the MCMC algorithm consists of sampling values of the added parameter λ, denoted as Bayesian Modification Indices (BMI), given the sampled values of the other IRT model parameters. These extra draws do not influence the chain, and the Markov chain remains restricted to the manifold of the posterior of the null-model. Theoretical results will show that the resulting estimate of the marginal posterior is a good approximation of the true marginal posterior distribution. As a result, various model violations can be tested by sampling extra parameters in the MCMC algorithm.

The outline of the paper is as follows. In Section 2, Bayesian modification (BMI) indices and the specific model violations are defined for IRT models. Then, in

© VVS, 2005

Section 3, the performance of the method is evaluated in a number of simulation studies.

2 Bayesian modification indices for IRT models

In this section, two BM indices will be presented. These indices are the modification indices for the item characteristic functions, local independence, and for differential item functioning (GLAS, 1998, 1999). Exact definitions of these model violations will be given below; a general framework will be outlined first.

The 2PNO model (LORD and NOVICK, 1968) will be considered as a null-model. From this null-model, a more general model is derived by adding parameters such that the assumption to be tested is violated. In the 2PNO model, the probability of a correct response of person i to item j, denoted $Y_{ij} = 1$, is given by

$$P(Y_{ij} = 1; \theta_i, \alpha_j, \beta_j) = \Phi(\eta_{ij}),$$

where Φ denotes the standard normal cumulative distribution function, and $\eta_{ij} = \alpha_j\theta_i - \beta_j$ with θ_i the ability of person i, α_j, the discrimination parameter and β_j the difficulty parameter of the item j, respectively.

In general, Bayesian modification indices for the 2PNO are derived as follows. Suppose that item j is the item of interest, and that the alternative model parameters are $\boldsymbol{\zeta}$ and $\boldsymbol{\delta}$. A general model will be defined as

$$\eta_{ij} = \alpha_j\theta_i - \beta_j + \mathbf{x}_i^t(\boldsymbol{\zeta}\theta_i - \boldsymbol{\delta}), \tag{1}$$

where $\mathbf{x}_i^t(\boldsymbol{\zeta}\theta_i - \boldsymbol{\delta})$ is the inner product of an explanatory variable $\mathbf{x}_i$ and a vector that is a function of the alternative model parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\delta}$. It must be noted that equation (1) is a general formulation, in many applications the number of alternative parameters per item may be one only. Obviously, interest is focused on the BMI values of the parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\delta}$.

An MCMC algorithm for generating the posterior distribution of the parameters of the 2PNO is described by Albert (1992). This implementation of the MCMC algorithm involves a data augmentation step which produces 'pseudo-data' Z_{ij} defined by

$$Z_{ij} = \alpha_j\theta_i - \beta_j + \epsilon_{ij},$$

where ϵ_{ij} is a normally distributed error variable. Within this MCMC algorithm for the null-model the definition of the augmented data Z_{ij} can be extended to

$$Z_{ij} = \alpha_j\theta_i - \beta_j + \mathbf{x}_i^t(\boldsymbol{\zeta}\theta_i - \boldsymbol{\delta}) + \epsilon_{ij}, \tag{2}$$

Notice that (2) implies a normal regression model with $Z_{ij} - \alpha_j\theta_i + \beta_j$ as dependent variables, $\mathbf{x}_i$ and θ_i as predictor variables and $\boldsymbol{\zeta}$ and $\boldsymbol{\delta}$ as regression coefficients. So, the full conditionals of $\boldsymbol{\zeta}$ and $\boldsymbol{\delta}$ follow from this linear regression model, and both parameters are easily sampled. Notice that (2) also provides a nice interpretation of

© VVS, 2005

the procedure: the magnitude of the draws of $\boldsymbol{\zeta}$ and $\boldsymbol{\delta}$ depend on the extent to which the difference between Z_{ij} and $\alpha_j\theta_i - \beta_j$ is properly modelled under the null model and the extent to which adding a predictor $\mathbf{x}_i$ can improve the null-model.

The extension of the null-model, equation (2), can always be written as a linear regression model with two components,

$$\mathbf{Z}_j = \mathbf{x}_1\boldsymbol{\xi}_j + \mathbf{x}_2\boldsymbol{\lambda} + \boldsymbol{\epsilon}_j, \tag{3}$$

where the parameters are defined as $\boldsymbol{\xi}_j = (\alpha_j, \beta_j)$ and $\boldsymbol{\lambda} = (\boldsymbol{\zeta}, \boldsymbol{\delta})$ and the non-observed augmented data are defined as $\mathbf{Z}_j = (Z_{1j},\ldots,Z_{nj})^t$. The covariates $\mathbf{X}$ are split up in two components $\mathbf{x_1}$ and $\mathbf{x_2}$, and contain the covariates corresponding to parameters $\boldsymbol{\xi}_j$ and $\boldsymbol{\lambda}$, respectively. Given the parameters $\mathbf{Z}_j$ and $\mathbf{X}$, the parameters $\boldsymbol{\xi}_j$ and $\boldsymbol{\lambda}$ can be estimated using least squares, that is

$$\begin{pmatrix} \hat{\boldsymbol{\xi}}_j \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^t\mathbf{x}_1 & \mathbf{x}_1^t\mathbf{x}_2 \\ \mathbf{x}_2^t\mathbf{x}_1 & \mathbf{x}_2^t\mathbf{x}_2 \end{pmatrix}^{-1} \mathbf{X}^t\mathbf{Z}_j = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \mathbf{X}^t\mathbf{Z}_j = \mathbf{V}\mathbf{X}^t\mathbf{Z}_j,$$

with $\mathbf{V} = (\mathbf{X}^t\mathbf{X})^{-1}$ and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)$. Note that interest is focused on the least squares estimate of $\boldsymbol{\lambda}$ for the computation of the BMI values. In the Appendix, the following theorem is proved.

THEOREM 1. *The marginal BM distribution of parameters $\boldsymbol{\lambda}$ of dimension r in equation (3) given the augmented data, employing a noninformative reference prior for the regression parameters and residual variance parameter, is a multivariate t-distribution,*

$$\boldsymbol{\lambda} \mid \mathbf{Z}_j \sim t_r\left[\hat{\boldsymbol{\lambda}}, s_0^2\mathbf{V}_{22}, n-2\right],$$

where $\mathbf{V}_{22}$ is the submatrix of $\mathbf{V}$ associated with $\hat{\boldsymbol{\lambda}}$, and s_0^2 is an estimate of the residual variance under the null-model.

This BM distribution resembles the true marginal posterior distribution except that the residual variance is based upon the null-model, and the increase of r degrees of freedom. With a sufficient number of observations, inferences from the BM distribution are useful due to the correspondence with the true marginal posterior distribution.

Suppose that the null-model holds. The alternative model parameters are all equal to zero, that is, $\boldsymbol{\lambda}_0 = \mathbf{0}$. Then the parameters $\boldsymbol{\lambda}_0$ are located inside the highest posterior density (HPD) region if and only if:

$$P[p(\boldsymbol{\lambda} \mid \mathbf{y}) > p(\boldsymbol{\lambda}_0 \mid \mathbf{y}) \mid \mathbf{y}] \leq 1 - \tilde{\alpha}, \tag{4}$$

where $1 - \tilde{\alpha}$ is some predefined credible level, say 0.90, and $p(\boldsymbol{\lambda} \mid \mathbf{y})$ is considered a random variable.

If the model is violated, the point $\boldsymbol{\lambda}_0$ is not captured by the HPD region, and this can be regarded as an indication of the existence of a model violation. One minus the content of the HPD region which just covers $\boldsymbol{\lambda}_0$ can be regarded as a Bayesian

© VVS, 2005

p-value (see BOX and TIAO (1973); HELD (2004)). It gives the significance level associated with the null hypothesis $\boldsymbol{\lambda}_0 = \mathbf{0}$ against the alternative $\boldsymbol{\lambda}_0 \neq \mathbf{0}$. It provides the posterior evidence against a given point based on the HPD region. There is a large amount of literature about different *p*-values and the philosophy behind it (see e.g. BAYARRI and BERGER (1999, 2000); GELMAN, CARLIN, STERN and RUBIN (1995); MENG (1994) and references therein). However, it is beyond the scope of the present article to discuss the use of different *p*-values.

THEOREM 2. *The Bayesian p-value of any point $\boldsymbol{\lambda}_0$ can be calculated based on the fact that the quantity*

$$(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}})^t \mathbf{V}_{22}^{-1} (\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}) / (r s_0^2)$$

is F-distributed with r and $n - 2$ degrees of freedom given the augmented data $\mathbf{Z}$. *In particular, a single parameter λ has the distribution*

$$\frac{\lambda - \hat{\lambda}}{s_0 \sqrt{V_{22}}} = t_{n-2}.$$

This result follows directly from the fact that $\boldsymbol{\lambda}$ follows a multivariate t-distribution. Marginal Bayesian *p*-value or HPD regions are computed by averaging over the sampled values of the augmented data, say $\mathbf{Z}^{(1)},\ldots,\mathbf{Z}^{(M)}$ with the use of an MCMC algorithm. For example, the computation of a Bayesian *p*-value follows from Theorem 2, that is,

$$\begin{aligned}\hat{\pi}(\boldsymbol{\lambda}_0) &= 1/M \sum_{m=1}^{M} P\left(\tilde{p}(\boldsymbol{\lambda} \mid \mathbf{Z}^{(m)}, \mathbf{y}) \leq \tilde{p}(\boldsymbol{\lambda}_0 \mid \mathbf{Z}^{(m)}, \mathbf{y}) \mid \mathbf{Z}^{(m)}, \mathbf{y}\right) \\ &= 1/M \sum_{m=1}^{M} P\left(F_{(r,n-2)} < \frac{\left(\boldsymbol{\lambda}_0 - \hat{\boldsymbol{\lambda}}\right)^t \mathbf{V}_{22}^{-1} \left(\boldsymbol{\lambda}_0 - \hat{\boldsymbol{\lambda}}\right)}{r s_0^2}\right),\end{aligned}$$

where $F_{(r,n-2)}$ is an *F*-variable with *r* and $n - 2$ degrees of freedom, and $\pi(\boldsymbol{\lambda}_0)$ the probability of an HPD just including $\boldsymbol{\lambda}_0$. A significant modification index can be seen as a caution index for further investigation. Besides the HPD region and Bayesian *p*-value, the mean and variance of the sampled values may provide a further indication of the possible importance of the violation.

2.1 Local independence

Here, $\mathbf{x}_i$ is a one-dimensional vector with an element y_{ij} and $\boldsymbol{\delta}$ a one-dimensional vector with an element δ_{jk} that models the dependence between the items *j* and *k*. In this model, which was originally proposed by KELDERMAN (1984) in the framework of the Rasch model, it is assumed that the magnitude of the dependence between the responses does not depend on the latent variable θ_i.

© VVS, 2005

This model, equation (2), is a linear regression model given the augmented $\mathbf{Z}_i$ values as a dependent variable and $\boldsymbol{\theta}$ and $\mathbf{y}_i$ as explanatory variables. That is,

$$Z_{ij} = \alpha_j\theta_i - \beta_j + y_{ij}\delta_{jk} + \epsilon_{ij}, \quad \text{for } i = 1, \ldots, n,$$

where the errors are independently normally distributed. Then, the BM distribution of δ_{jk} given the augmented data is a t-distribution. This BM distribution can be used to test the null-hypothesis $\delta_{jk} = 0$.

2.2 Differential item functioning

Differential item functioning (DIF) is a difference in item responses between equally proficient members of two or more groups. Usually, one distinguishes a reference group, say the majority population, and one or more focal groups, say disadvantaged groups. A dichotomous item is subject to DIF if, conditionally on proficiency level, the probability of a correct response differs between groups. One might think of a test of foreign language comprehension, where girls are impeded by items referring to a football setting. The poor performance of the girls on the football-related items must not be attributed to their poor level of comprehension of the foreign language but to their lack of knowledge of football. Since DIF is highly undesirable in fair testing, methods for detection of DIF are extensively studied (see, for instance, HOLLAND and WAINER (1993) or CAMILLI and SHEPARD (1994)) and various methods for detection of DIF have been proposed (HOLLAND and THAYER, 1988; HAMBLETON and RODGERS, 1989; KELDERMAN, 1989; SWAMINATHAN and RODGERS, 1990; MURAKI and BOCK, 1991).

The definition of a Bayesian modification for DIF proceeds as follows. Define a background variable

$$x_i = \begin{cases} 1 & \text{if } i \text{ belongs to the focal group,} \\ 0 & \text{if } i \text{ belongs to the reference group.} \end{cases}$$

Usually, two forms of DIF are distinguished: uniform DIF, where the difference of the probability of a correct response between groups does not depend on the value of the latent trait, and non-uniform DIF, where interaction between this difference and the latent trait does exist (see MELLENBERGH (1982, 1983)). In the present framework, this can be modelled by adding parameters δ_j for modelling uniform DIF and δ_j and ζ_j for modelling non-uniform DIF, respectively.

3 Some power studies

A number of simulation studies were conducted to investigate the power of the procedure proposed here. The aim of the first simulation presented here is to illustrate the data used in the power studies. The data were generated under the 2PNO model. The item parameter values are given in the first columns of the first panel of Table 1. The simulees were drawn from two groups. These groups had

© VVS, 2005

Table 1. Simulation values and data summary. Number of observations equals 2000.

			p-value			Frequency	
Item	α_i	β_i	$g = 1$	$g = 2$	Score	$g = 1$	$g = 2$
					0	32	11
1	0.40	−1.00	0.41	0.84	1	64	26
2	0.60	−1.00	0.40	0.83	2	90	54
3	0.80	−1.00	0.39	0.83	3	132	103
4	0.40	0.00	0.24	0.50	4	155	152
5	0.60	0.00	0.26	0.55	5	164	190
6	0.80	0.00	0.24	0.55	6	156	194
7	0.40	1.00	0.9	0.20	7	114	145
8	0.60	1.00	0.11	0.24	8	63	84
9	0.80	1.00	0.10	0.24	9	23	35
10	0.60	0.00	0.25	0.56	10	0	0

Group	μ_θ	σ_θ	Mean	SD	Alpha
1	0.00	1.0	4.6	2.2	0.58
2	0.50	0.8	5.2	2.0	0.49

different normal ability distributions with means and standard deviations as displayed in the first columns of the second panel of Table 1. Every group consisted of 1000 simulees. An example of classical statistics of a simulated data set are displayed in the last columns of the two panels of Table 1. The columns of the first panel contain the true item parameters, and the *p*-values and the distributions of the test takers' sum scores, for the two groups, respectively. The columns of the second panel contain the mean and the standard deviation of the distributions of the ability parameters per group, the mean and the standard deviation of the score distributions per group, and the coefficient alpha per group.

After computing the posterior distributions of the item and population parameters using MCMC, parameter estimates and credible intervals were computed as the mean and standard deviation of the posterior distributions, respectively. These estimates are given in Table 2. The number of MCMC iterations was 1000 for the burn-in period and 3000 for generating the actual estimates. Notice that the model is identified by setting the mean and the standard deviation of the ability distribution of the first group equal to zero and one, respectively.

In Table 2, the last two columns give the significance probabilities for the Bayesian modification indices targeted at DIF and violation of local independence, respectively. The BMI targeted at local independence was used to test the dependence of the item response on item j on the response to the previous item $j - 1$. Therefore, in the last column there is no significance probability for the first item.

Using the generating values of the item and population parameters of Table 1, a number of power studies were performed. These studies concerned the Type I error rate under the null-model, the hit rate, that is, the power to detect model violations, and the false alarm rate, that is, the Type I error rate for model-conform items in a test where one or more of the other items violated the model. The studies on the

© VVS, 2005

Table 2. Estimated parameter values, standard deviations and fit indices. Number of observations equals 2000.

Item	α_i	sd(α_i)	β_i	sd(β_i)	BMI_{Dif}	BMI_{Loc}
1	0.57	0.08	−1.01	0.06	0.26	–
2	0.63	0.08	−1.02	0.06	0.58	0.54
3	0.82	0.10	−1.07	0.07	0.46	0.58
4	0.42	0.06	0.06	0.04	0.44	0.33
5	0.62	0.07	−0.04	0.05	0.55	0.44
6	1.05	0.13	0.03	0.07	0.39	0.52
7	0.36	0.07	0.95	0.05	0.43	0.61
8	0.60	0.08	0.91	0.06	0.49	0.19
9	0.96	0.14	1.09	0.11	0.66	0.71
10	0.58	0.07	−0.06	0.05	0.61	0.48

Group	μ_θ	sd(μ_θ)	σ_θ	sd(σ_θ)
1	0.00	–	1.00	–
2	0.37	0.07	0.78	0.07

Type I error rate under the null-model were conducted for sample sizes of 1000 and 2000 simulees. Further, apart from a test length of 10 items, a test of 20 items was generated by duplicating the true parameter values of Table 1. Model violations were simulated by generating responses using a non-zero value for a δ-parameter. Introducing non-zero ζ-parameters is beyond the scope of this study. The model violation was always imposed on the last item of the test. For all studies reported below, 100 replications were made. In every replication, the probability was computed that the parameter point $\delta = 0$ was covered by the 90% HPD region, according to equation (4), using the sampled parameter values from the BM distribution. Further, the Bayesian p-values associated with the null hypothesis $\delta = 0$ against the alternative $\delta \neq 0$, were computed. Power was defined as the percentage of significantly shifted BM distributions over replications.

Two series of simulations were performed, the first one aimed at DIF, and the second at violation of local independence.

The first series of simulation studies entailed the power of the test based on the BM for DIF. Data were generated using the true item and population parameters for Table 1, with the distinction that a DIF parameter $\delta = 0.10$, $\delta = 0.20$, or $\delta = 0.50$ was added to the last item. The results are shown in Table 3. The reported sample sizes refer to the number of simulees in each of the two groups. As expected, the hit rate is an increasing function of both the effect size, the sample size and the number of items. Not shown in the table is the fact that for all combinations of sample size and test length, the false alarm rate of all BM tests was approximately 10%.

In the second series of simulations, the power of the BM distribution targeted at violation of local independence was investigated. In these simulations, only one group of test takers with a standard normal ability distribution was used. Further, the item parameters were as displayed in Table 1, with the distinction that a parameter

© VVS, 2005

Table 3. Power of the BM for DIF and local independence (LOC), detection percentage.

		DIF			LOC		
		Effect Size			Effect Size		
N	K	0.10	0.20	0.50	0.10	0.20	0.50
1000	10	16	78	100	0	4	51
	20	24	85	100	0	5	77
2000	10	33	94	100	0	4	83
	20	34	98	100	0	3	93

$\delta = 0.10$, $\delta = 0.20$, or $\delta = 0.50$ was added to model the dependency between item 9 and 10. The BM distributions were computed for every consecutive pair of items. The results for the BM distributions of the last item are shown in Table 3 under the label LOC. The results are generally analogous to the results of DIF, with the exception that the power for the combination of the smallest sample and effect size becomes negligible. The false alarm rate of the BM distributions for the other items and the other model violation considered here were close to the nominal significance probability.

4 Discussion

In this paper, it was investigated how violations to multi-parameter models can be evaluated in a Bayesian framework. The main advantage of the approach is that many model violations for all items can be assessed without complicated and time consuming computations. Analogous to modification indices in a frequentist framework, modification indices mainly serve a purpose as caution indices, and a significant result can be followed by a more detailed analysis with more traditional tools such as Bayes factor, posterior predictive checks, or the Bayesian Information Criterion (BIC).

A point of further study is the generalization of the approach to more complex models. One of the main advantages of estimating IRT models using a fully Bayesian approach is that traditional frequentist approaches break down because of the infeasible numerical evaluation of the multiple integrals involved in solving the estimation equations. In the framework of computer adaptive testing, interesting models are testlet response models (Bradlow, Wainer and Wang, 1999), models with multidimensional latent abilities (Béguin and Glas, 2001) and multilevel IRT models (Fox and Glas, 2001, 2003) and it is in the realm of these models that more research needs to be done.

Appendix: Proof of Theorem 1

According to equation (3), define

© VVS, 2005

$$\mathbf{V}^{-1} = \left([\mathbf{x}_1, \mathbf{x}_2]^t[\mathbf{x}_1, \mathbf{x}_2]\right) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}^{-1} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix},$$

where $\mathbf{x}_1$ is an $n \times 2$ and $\mathbf{x}_2$ an $n \times r$ matrix with rank 2 and r, respectively. Under the null-model, $\boldsymbol{\lambda} = 0$, the conditional distribution of the parameters of item j equals

$$\boldsymbol{\xi}_j \mid \mathbf{Z}, \mathbf{x}_1, \sigma_0^2 \sim \mathcal{N}\left(\hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}, \sigma_0^2 V_{11}\right)$$

where $\hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}$ is the least-squares estimate of $\boldsymbol{\xi}_j$, given $\mathbf{x}_1$ and $\mathbf{Z}$. It is assumed that the elements of $\boldsymbol{\lambda}$, $\boldsymbol{\xi}_j$ and $\log(\sigma_0^2)$ are uniformly and independently distributed. The BMI values of $\boldsymbol{\lambda}$ are sampled from its full conditional distribution, that is,

$$\boldsymbol{\lambda} \mid \mathbf{Z}, \mathbf{x}, \boldsymbol{\xi}_j, \sigma_0^2 \sim \mathcal{N}\left(\hat{\boldsymbol{\lambda}} + V_{12}^t V_{11}^{-1}\left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}\right), \sigma_0^2 W_{22}^{-1}\right),$$

where $\hat{\boldsymbol{\lambda}}$ is the least-squares estimate of $\boldsymbol{\lambda}$ given $\mathbf{x}$ and $\mathbf{Z}$. The BM distribution can be obtained by integration,

$$\begin{aligned}
\tilde{p}(\boldsymbol{\lambda} \mid \mathbf{x}, \mathbf{Z}) &= \int\int \tilde{p}(\boldsymbol{\lambda} \mid \mathbf{Z}, \mathbf{x}, \boldsymbol{\xi}_j, \sigma_0^2) p(\boldsymbol{\xi}_j, \sigma_0^2 \mid \mathbf{Z}, \mathbf{x}_1) \mathrm{d}\boldsymbol{\xi}_j \, \mathrm{d}\sigma_0^2 \\
&\propto \int\int \sigma_0^{-(n+r+1)} \exp\left(\frac{-1}{2\sigma_0^2}\left(\left[\left(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\right) - V_{12}^t V_{11}^{-1}\left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}\right)\right]^t W_{22}\right.\right. \\
&\quad \times \left[\left(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\right) - V_{12}^t V_{11}^{-1}\left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}\right)\right] + (n-2)s_0^2 \\
&\quad \left.\left. + \left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}\right)^t V_{11}^{-1}\left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}\right)\right)\right) \mathrm{d}\boldsymbol{\xi}_j \, \mathrm{d}\sigma_0^2. \qquad (5)
\end{aligned}$$

The expression within the exponent can be recognized as a specific form of the inverse of the partitioned full rank symmetric matrix $\mathbf{V}$, see, for example, SEARLE (1971, p. 27), that is,

$$\left[\left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_{j,\mathbf{x}_1}\right)^t, \left(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\right)^t\right] \begin{bmatrix} V_{11}^{-1} + V_{11}^{-1} V_{12} W_{22} V_{12}^t V_{11}^{-1} & -V_{11}^{-1} V_{12} W_{22} \\ -W_{22} V_{12}^t V_{11}^{-1} & W_{22} \end{bmatrix} \begin{bmatrix} \left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_{j,\mathbf{x}}\right) \\ \left(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\right) \end{bmatrix}.$$

Since this term represents the inverse of matrix $\mathbf{V}$, equation (5) with $\boldsymbol{\Lambda} = (\boldsymbol{\xi}_j, \boldsymbol{\lambda})$ simplifies to

$$\begin{aligned}
\tilde{p}(\boldsymbol{\Lambda} \mid \mathbf{Z}) &\propto \int\int \sigma_0^{-(n+r+1)} \exp\left(\frac{-1}{2\sigma_0^2}((n-2)s_0^2 + (\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}})^t \mathbf{V}^{-1} (\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}))\right) \mathrm{d}\boldsymbol{\xi}_j \, \mathrm{d}\sigma_0^2 \\
&\propto \int \left[v s_0^2 + \left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)^t \mathbf{V}^{-1} \left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)\right]^{-(v+2+r)/2} \mathrm{d}\boldsymbol{\xi}_j, \qquad (6)
\end{aligned}$$

where $v = n - 2$. The integrand in (6) is in the form of a multivariate t-distribution. As a result, the marginal distribution of a r-dimensional subset has the multivariate t-distribution (BOX and TIAO, 1973)

$$\boldsymbol{\lambda} \mid \mathbf{Z}_j \sim t_r\left[\hat{\boldsymbol{\lambda}}, s_0^2 V_{22}, v\right].$$

© VVS, 2005

References

AITCHISON, J. and D. S. SILVEY (1958), Maximum likelihood estimation of parameters subject to restraints, *Annals of Mathematical Statistics* **29**, 813–828.

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

BAYARRI, M. J. and J. O. BERGER (1999), Quantifying suprise in the data and model verification, in: J. M. BERNARDO, J. O. BERGER, A. P. DAWID and A. F. M. SMITH (eds), *Bayesian statistics* **6**, 53–82, Oxford University Press, London.

BAYARRI, M. J. and J. O. BERGER (2000), P values for composite null models, *Journal of the American Statistical Association* **95**, 1127–1142.

BERNARDO, J. M. and A. F. M. SMITH (1994), *Bayesian theory*, John Wiley & Sons, New York.

BÉGUIN, A. A. and C. A. W. GLAS (2001), MCMC estimation of multidimensional IRT models, *Psychometrika* **66**, 541–561.

BOX, G. E. P. and G. C. TIAO (1973), *Bayesian inference in statistical analysis*, Addison-Wesley, Reading, MA.

BRADLOW, E. T., H. WAINER and X. WANG (1999), A Bayesian random effects model for testlets, *Psychometrika* **64**, 153–168.

CAMILLI, G. and L. A. SHEPARD (1994), *Methods for identifying biased test items*, Thousand Oaks, Sage, CA.

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

FOX, J.-P. and C. A. W. GLAS (2003), Bayesian modeling of measurement error in predictor variables using item response theory, *Psychometrika* **68**, 169–191.

GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN (1995), *Bayesian data analysis*, Chapman & Hall, London.

GLAS, C. A. W. (1998), Detection of differential item functioning using Lagrange multiplier tests, *Statistica Sinica* **8**, 647–667.

GLAS, C. A. W. (1999), Modification indices for the 2-pl and the nominal response model, *Psychometrika* **64**, 273–294.

HAMBLETON, R. K. and H. J. RODGERS (1989), Detecting potentially biased test items: comparison of IRT area and Mantel–Haenszel methods, *Applied Measurement in Education* **2**, 313–334.

HELD, L. (2004), Simultaneous posterior probability statements from Monte Carlo output, *Journal of Computational and Graphical Statistics* **13**, 20–35.

HOLLAND, P. W. and H. WAINER (1993), *Differential item functioning*, Erlbaum, Hillsdale, NJ.

HOLLAND, P. W. and D. T. THAYER (1988), Differential item functioning and the Mantel–Haenszel procedure, in: H. WAINER and H. I. BRAUN (eds), *Test validity*, Lawrence Erlbaum Associates Inc, Hillsdale, NJ.

KELDERMAN, H. (1984), Loglinear RM tests, *Psychometrika* **49**, 223–245.

KELDERMAN, H. (1989), Item bias detection using loglinear IRT, *Psychometrika* **54**, 681–697.

LINDLEY, D. V. (1965), *Introduction to probability and statistics from a Bayesian viewpoint* (2 vols - *Part I: Probability* and *Part II: Inference)*, Cambridge University Press, Cambridge.

LORD, F. M. and M. R. NOVICK (1968), *Statistical theories of mental test scores*, Addison-Wesley, Reading, MA.

MELLENBERGH, G. J. (1982), Contingency table models for assessing item bias, *Journal of Educational Statistics* **7**, 105–118.

MELLENBERGH, G. J. (1983), Conditional item bias methods, in: S. H. IRVINE and W. J. BERRY (eds), *Human assessment and cultural factors*, Plenum Press, New York.

MENG, X. L. (1994), Posterior predictive p-values, *The Annals of Statistics* **22**, 1142–1160.

MURAKI, E. and R. D. BOCK (1991), *PARSCALE: Parameter scaling of rating data* [computer program], Scientific Software, Inc., Chicago, IL.

© VVS, 2005

RAO, C. R. (1947), Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

SEARLE, S. R. (1971), *Linear models*, Wiley, New York.

SWAMINATHAN, H. and H. J. RODGERS (1990), Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement* **27**, 361–370.

ZELLNER, A. (1971), *An introduction to Bayesian inference in econometrics*, Addison-Wesley, Reading, MA.

Received: June 2004. Revised: December 2004.

© VVS, 2005