

■ **A Bayesian Approach to Item Calibration and
Evaluation of Parameter Drift**

Cees A. W. Glas
University of Twente, Enschede, The Netherlands

■ **Law School Admission Council**
Computerized Testing Report 00-02
December 2005

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

© 2005 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any part or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, 662 Penn Street, Box 40, Newtown, PA 18940-0040.

LSAT and LSAC are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary	1
Abstract	1
Introduction	1
Bayesian Estimation of the 3-PNO via MCMC.	2
Modification Indices in a Bayesian Framework	5
Bayesian Modification Indices for the 3-PNO Model	6
<i>Local Independence</i>	6
<i>Differential Item Functioning and Parameter Drift</i>	7
<i>The Form of the ICCs</i>	7
A Power Study.	8
Discussion	10
References.	10

Executive Summary

In previous reports for the Law School Admission Council (LSAC), methods for evaluating possible differences between item parameter values during pretesting and the operational stage of a Computerized Adaptive Testing (CAT) program were investigated. These differences are often labeled *parameter drift* and can have various causes. For example, frequent exposure of items may affect their difficulty, as may motivational differences between examinees in the pretest and the operational test.

Earlier tests to detect item parameter drift proposed by the author were based on marginal maximum likelihood (MML) estimation of the values of the item parameters. Recently, however, a Bayesian approach has been proposed as an alternative to the MML framework, which allows for items with a more complicated response format. Applications include items with multiple raters, testlet structures, multidimensional latent abilities, and multilevel structures for the ability parameter.

In the present paper, the evaluation of parameter drift within this Bayesian framework was investigated. A number of Bayesian modification (BM) indices for the three-parameter normal ogive model are presented, which are generalizations of the modification indices proposed by the author for the MML framework. The strong point of the BM procedure is that many model violations for all items can be assessed without complicated, time-consuming computations.

The power of these indices was investigated in a series of computer simulations. As expected, the power was an increasing function both of the size of the examinee samples and the effect size (true amount of item parameter drift). Generally, the power was smaller than the power of the earlier procedure for the MML framework. However, the likelihood of Type 1 errors (false alarms) was virtually equal to zero. It was therefore concluded that the proposed modification indices could serve very well as quick and convenient caution indices, with significant results followed by a more detailed traditional analysis.

Abstract

In previous reports for the Law School Admission Council (LSAC), methods for evaluating differences between the parameter estimates of the pretest and online phase were evaluated. These differences are often labeled *parameter drift*. Three approaches were suggested, based on a Lagrange Multiplier (LM) statistic, a Wald statistic and a cumulative sum (CUSUM) statistic, respectively. All three methods are based on a marginal maximum likelihood (MML) framework. Recently, however, an alternative to the MML framework has been proposed, which is based on a Bayesian approach. In this approach, a Markov chain Monte Carlo (MCMC) procedure is used to construct posterior distributions of the parameters of interest. In the present paper, methods for how to evaluate parameter drift; that is, differences between the parameter estimates of the pretest and online phase in this Bayesian framework were investigated. First, a Bayesian procedure to estimate the three-parameter normal ogive model is proposed. Then, a number of Bayesian modification indices are presented that can be viewed as generalizations of modification indices proposed for the MML framework. Finally, the performance of the method is evaluated in a number of simulation studies.

Introduction

Recently, Glas and Verhelst (1995, also see, Glas, 1998, 1999, 2001, 2003) proposed using modification indices for evaluating model fit to item response theory (IRT) models. These modification indices are based on the Lagrange Multiplier (LM) Test (Aitchison & Silvey, 1958), and the Equivalent Efficient Score Test (Rao, 1947). Modification indices are targeted at specific model assumptions, such as the form of the item characteristic curves (ICCs), the form of the ability distribution, local independence, and item parameter invariance. Apart from producing significance probabilities, the procedure also provides indices helpful in assessing the severity of model violations. In this approach, model violations are assessed by constructing a more general model where assumptions are relaxed by adding parameters. Besides being targeted at specific model violations, modification indices are also item oriented. So the procedure is aimed at producing a wide range of diagnostics that can be used for item selection or for adjusting the model. Therefore, estimating all possible alternative models is no option, and the procedure is based on the parameters of the null-model, that is, the item response theory model, only. These estimates can be obtained by marginal maximum likelihood (MML) (Bock & Aitkin, 1981) or by a Bayes modal procedure (Mislevy, 1986).

Formally, the LM test is used to test a special model against a general model, where the special model is derived from the general model by imposing constraints on the parameter space. In many instances, this is accomplished by setting one or more parameters of the general model equal to zero. The LM test is based on evaluation of the first-order partial derivatives of the log-likelihood function of the general model, evaluated using the ML estimates of the restricted model. The unrestricted elements of the vector of first order derivatives are equal to zero, because their values originate from solving the likelihood equations. The

magnitude of the elements of the vector of first-order partial derivatives corresponding to the restricted parameters determine the value of the statistic: the closer they are to zero, the better the model fit.

As an example, consider a test for local independence for the two-parameter logistic (2PL) model proposed by Glas (1999). In the 2PL model, the probability of a correct response of person i to item j , denoted $Y_{ij} = 1$, is given by

$$P(Y_{ij} = 1 | \theta_i, \alpha_j, \beta_j) = \frac{\exp(a_j \theta_i - \beta_j)}{1 + \exp(a_j \theta_i - \beta_j)}, \quad (1)$$

where θ_i , α_j , and β_j are the ability, item discrimination, and item difficulty parameter, respectively. Suppose that the local independence between this item and some item k is violated. Then, following Kelderman (1984, also see, Jannarone, 1986) the probability of a correct response to item k can be given by

$$P(Y_{ik} = 1 | Y_{ij} = y_{ij}, \theta_i, \alpha_j, \beta_j) = \frac{\exp(a_k \theta_i - \beta_k - y_{ij} \delta_{jk})}{1 + \exp(a_k \theta_i - \beta_k - y_{ij} \delta_{jk})}, \quad (2)$$

where δ_{jk} models the dependence between the items j and k . In the special model, the 2PL model, $\delta_{jk} = 0$. The LM test is based on the derivative with respect to δ_{jk} evaluated in $\delta_{jk} = 0$. Further, the importance of a possible significant result, say the effect size, can be evaluated by computing a one-Newton-Raphson step estimate of the parameter. This one step estimate can be directly obtained from the LM statistic (Glas, 1999).

The purpose of this article is to develop an analogous procedure for fully Bayesian estimation procedures for IRT models, keeping intact its most salient features. So the aim is to develop a diagnostic, item-oriented tool aimed at specific model violations that is based on estimates issued from the null-model, and which provides significance probabilities and information on the importance of model violations. A fully Bayesian framework for estimation of IRT models was proposed by Albert (1992). From a theoretical point of view, the fully Bayesian framework applied to the 2PL model may have some advantages over the MML framework (see, for instance, Baker, 1998), but these advantages may not be the most important reasons for its adoption. More importantly, the fully Bayesian framework results in a straightforward and easily implemented estimation routine (Patz & Junker, 1999) that is easily generalized to more complex models. Recently, this framework has been adopted for the estimation of IRT models with multiple raters (Patz & Junker, 1997), testlet structures (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, and Du, 1999), latent classes (Hojtink & Molenaar, 1997), multidimensional latent abilities (Béguin & Glas, 1998) and linear multilevel models on the ability parameters (Fox & Glas, 1998). In the present article, Bayesian modification indices will be described in detail for the three-parameter model. In principle, however, the approach is applicable to all the complex IRT models referenced.

This paper is organized as follows: First, the Bayesian estimation procedure for the three-parameters normal ogive (3-PNO) model and a testlet response model based on the 3-PNO will be sketched. Then, an explanation of how Bayesian modification indices can be implemented in this framework will be provided. Finally, some power studies will be presented.

Bayesian Estimation of the 3-PNO via MCMC

In this paper, a Markov chain Monte Carlo (MCMC) procedure will be used to generate the posterior distributions of interest. The MCMC chains will be constructed using the Gibbs sampler (Gelfand & Smith, 1990). To implement the Gibbs sampler, the parameter vector is divided into a number of components, and each successive component is sampled from its conditional distribution given sampled values for all other components. This sampling scheme is repeated until the sampled values form stable posterior distributions. Albert (1992) applies Gibbs sampling to estimate the parameters of the well known two-parameter normal ogive model (Lord & Novick, 1968). In this section, the procedure will be generalized to the 3-PNO. The 3-PNO is given by

$$\begin{aligned} P(Y_{ij} = 1; \theta_i, \alpha_j, \beta_j, \gamma_j) &= \gamma_j + (1 - \gamma_j) \Phi(\eta_{ij}) \\ &= \Phi(\eta_{ij}) + \gamma_j (1 - \Phi(\eta_{ij})), \end{aligned} \quad (3)$$

where γ_j is called the *pseudo-guessing parameter*, 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 the second line of equation (3), the usual expression for the 3-PNO is rewritten to an expression that supports an interpretation of the model that will be used below. In

this interpretation, there is a probability, $\Phi(\eta_{ij})$, that the respondent knows the item and gives a correct response with probability one, and a probability, $(1 - \Phi(\eta_{ij}))$, that the respondent does not know the item and guesses with γ_j as the probability of a correct response. So the probability of a correct response is a sum of a term $\Phi(\eta_{ij})$ and a term $\gamma_j(1 - \Phi(\eta_{ij}))$. In line with this interpretation, it will prove convenient to introduce a vector of binary variables W_{ij} such that

$$W_{ij} = \begin{cases} 1 & \text{if person } i \text{ knows the correct answer to item } j \\ 0 & \text{if person } i \text{ does not know the correct answer to item } j \end{cases} \quad (4)$$

So if $W_{ij} = 0$, person i guessed the response to item j , if $W_{ij} = 1$, person i knows the right answer and gives a correct response. Summing up

$$\begin{aligned} P(Y_{ij} = 1 | W_{ij} = 1, \eta_{ij}, \gamma_j) &= \Phi(\eta_{ij}) \\ P(Y_{ij} = 0 | W_{ij} = 1, \eta_{ij}, \gamma_j) &= 0 \\ P(Y_{ij} = 1 | W_{ij} = 0, \eta_{ij}, \gamma_j) &= \gamma_j \\ P(Y_{ij} = 0 | W_{ij} = 0, \eta_{ij}, \gamma_j) &= 1 - \gamma_j \end{aligned} \quad (5)$$

The estimation procedure will be sketched for data in a complete design, where all test takers respond to all items first. The generalization to incomplete data, such as the data issued from pretests for CAT and online data issued from CAT, will be treated at the end of this section.

Consider a design with G groups of test takers from G different populations. Assume that each group of test takers has a separate ability distribution. Let μ_g and σ_g^2 be the mean and variance of group g , $g = 1, \dots, G$ and let n_g be the number of test takers in group g . The model will be identified by choosing $\mu_1 = 0$ and $\sigma_1 = 1$.

To implement the Gibbs sampler, the data \mathbf{Y} , which are the responses of n test takers to k items, will be augmented with latent data $\mathbf{W} = (W_{11}, \dots, W_{nk})$. Following Albert (1992), the data are also augmented with latent data $\mathbf{Z} = (Z_{11}, \dots, Z_{nk})$, where the variables Z_{ij} are independent and normally distributed with mean η_{ij} and a standard deviation equal to one. These variables are related to \mathbf{W} by $Z_{ij} > 0$ if $W_{ij} = 1$ and $Z_{ij} \leq 0$ if $W_{ij} = 0$. This can be written as

$$p(Z_{ij} | W_{ij}, \eta_{ij}) \propto \phi(Z_{ij}; \eta_{ij}, 1) (\mathbf{I}(Z_{ij} > 0) \mathbf{I}(W_{ij} = 1) + \mathbf{I}(Z_{ij} \leq 0) \mathbf{I}(W_{ij} = 0)), \quad (6)$$

where $\phi(\cdot; \eta_{ij}, 1)$ stands for the normal density with mean η_{ij} and standard deviation equal to one. The item parameters ξ have a prior $p(\xi) = \prod_{j=1}^k \mathbf{I}(\alpha_j > 0)$, which ensures that the discrimination parameters are positive. The guessing parameter γ_j has the conjugate prior Beta(a, b).

The aim of the procedure is to simulate samples from the joint posterior distribution of ξ, θ, \mathbf{Z} , and \mathbf{W} , given by

$$\begin{aligned} p(\xi, \theta, \mathbf{Z}, \mathbf{W}, \mu, \sigma | \mathbf{Y}) &= p(\mathbf{Z}, \mathbf{W} | \mathbf{y}, \mathbf{D}; \xi, \gamma, \theta) p(\theta) p(\xi) p(\gamma) \\ &= C \prod_{i=1}^n \left\{ \prod_{j=1}^k p(Z_{ij} | W_{ij}, \eta_{ij}) p(W_{ij} | y_{ij}, \eta_{ij}, \gamma_j) \right\} \\ &\quad \phi(\theta_i; \mu_{g(i)}, \sigma_{g(i)}) (\mathbf{I}(\alpha_j > 0)) \prod_{j=1}^k p(\gamma_j) \end{aligned} \quad (7)$$

where $p(W_{ij} | y_{ij}, \eta_{ij}, \gamma_j)$ is given by equation (10) and $p(Z_{ij} | W_{ij}, \eta_{ij})$ follows from equation (2).

Although the distribution given by equation (7) has an intractable form, the fully conditional distributions of \mathbf{Z}, θ, ξ , and γ are each tractable and easy to sample from. A draw from the full conditional distribution can be obtained in the following six steps.

Step 1. Draw from the conditional distribution of Z_{ij} , given all other variables. Using equation (7), the distribution of Z_{ij} conditional on \mathbf{W}, θ , and ξ is given by

$$Z_{ij} | \mathbf{W}, \theta, \xi, \mathbf{y} \sim \begin{cases} N(\eta_{ij}, 1) \text{ truncated at the left by } 0 & \text{if } W_{ij} = 1 \\ N(\eta_{ij}, 1) \text{ truncated at the right by } 0 & \text{if } W_{ij} = 0 \end{cases}$$

Step 2. Draw from the conditional distribution of θ given \mathbf{Z} and ξ . From equations (2) and (7) it follows that $Z_{ij} - \beta_j = \alpha_j \theta_i + \varepsilon_{ij}$, where ε_{ij} is a normally distributed error term. So the full conditional distribution of θ entails a normal model for the regression of $Z_{ij} - \beta_j$ on α_j , with θ_i as a regression coefficient that has a normal prior with parameters $\mu_{g(i)}$ and $\sigma_{g(i)}$, where $g(i)$ is the group to which test taker i belongs. Therefore, the posterior of θ_i is normal, that is,

$$\theta_i \sim N\left(\frac{\hat{\theta}_i / v + \mu_{g(i)} / \sigma_{g(i)}^2}{1/v + 1/\sigma_{g(i)}^2}, \frac{1}{(1/v + 1/\sigma_{g(i)}^2)}\right), \quad (8)$$

where $\hat{\theta}_i = \sum_j \alpha_j (Z_{ij} + \beta_j) / \sum_j \alpha_j^2$ and $v = 1 / \sum_j \alpha_j^2$.

Step 3. Draw from the conditional distribution of the parameters of item j , $\xi_j = (\alpha_j, \beta_j)^T$. The variables ξ_j can be viewed as coefficients of the regression of $Z_j = (Z_{1j}, \dots, Z_{nj})^T$ on $\mathbf{X} = (\theta, -1)$, with -1 being the n dimensional column vector with elements -1 . Therefore,

$$\xi_j \mid \theta, \mathbf{Z}_j, \mathbf{y} \sim N(\hat{\xi}_j, (\mathbf{X}^T \mathbf{X})^{-1}) \mathbf{I}(\alpha_j > 0), \quad (9)$$

where $\hat{\xi}_j = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Z}_j$.

Step 4. Draw from the distribution of \mathbf{W} conditional on all other variables. From equation (5) it follows that the conditional probability of W_{ij} given Y_{ij} is given by

$$\begin{aligned} P(W_{ij} = 1 \mid Y_{ij} = 1, \eta_{ij}, \gamma_j) &\propto \Phi(\eta_{ij}) \\ P(W_{ij} = 0 \mid Y_{ij} = 1, \eta_{ij}, \gamma_j) &\propto \gamma_j (1 - \Phi(\eta_{ij})) \\ P(W_{ij} = 1 \mid Y_{ij} = 0, \eta_{ij}, \gamma_j) &= 0 \\ P(W_{ij} = 0 \mid Y_{ij} = 0, \eta_{ij}, \gamma_j) &= 1 \end{aligned} \quad (10)$$

Step 5. Sample from the conditional distribution of γ_j . Let t_j be defined as $t_j = \sum_{i=1}^n I(W_{ij} = 0)$, that is,

t_j is the number of test takers who do not know the correct answer to item j and guess the response. The probability of a correct response of test taker i on item j given $W_{ij} = 0$ is $P(y_{ij} = 1 \mid W_{ij} = 0) = \gamma_j$. The number of correct responses obtained by guessing, for example,

$s_j = \sum_{i \mid w_{ij}=0}^n y_{ij}$, has a binomial distribution, $\text{Bin}(t_j, \gamma_j)$.

With the noninformative conjugate Beta prior, the posterior distribution of γ_j is

$$\gamma_j \mid \mathbf{W}, \mathbf{y} \sim \text{Beta}(s_j + 1, t_j - s_j + 1). \quad (11)$$

Step 6. The conjugate prior distribution for (μ, σ) is a product of a normal and an inverse $-\chi^2$ distribution (see, for instance, Box & Tiao, 1973, or Gelman, Carlin, Stern, & Rubin, 1995). Since the model was identified choosing $\mu_1 = 0$ and $\sigma_1 = 1$, only priors for the ability distributions $g = 2, \dots, G$ will be specified. So for the parameters of these distributions, the priors are

$$\begin{aligned} \sigma_g^2 &\sim \text{Inverse } -\chi^2(v_0, \sigma_0^2) \\ \mu_g \mid \sigma_g, \theta_g &\sim N(\mu_0, \sigma_g^2 / \kappa_0), \end{aligned} \quad (12)$$

for $g = 2, \dots, G$. Then, the posterior is also a product of a normal and an inverse $-\chi^2$ distribution, that is

$$\begin{aligned} \sigma_g^2 &\sim \text{Inverse } -\chi^2(v_{gn}, \sigma_{gn}^2) \\ \mu_g \mid \sigma_g, \theta_g &\sim N(\mu_n, \sigma_{gn}^2 / \kappa_{gn}), \end{aligned} \quad (13)$$

where

$$\mu_{gn} = \frac{\kappa_0 n_g}{\kappa_0 + n_g} \mu_0 + \frac{n_g}{\kappa_0 + n_g} \bar{\theta}$$

$$v_{gn} = v_0 + n_g$$

$$\kappa_{gn} = \kappa_0 + n_g$$

$$v_{gn} \sigma_{gn}^2 = v_0 \sigma_0^2 + (n_g - 1) s_g + \frac{\kappa_0 n_g}{\kappa_0 + n_g} (\bar{\theta}_g - \mu_0)^2,$$

with $s_g = \sum_{i=1}^{n_g} (\theta_i - \bar{\theta}_g)(\theta_i - \bar{\theta}_g)$ and $\bar{\theta}_g$

the mean of the ability parameters in group g . In many situations, there are no a priori grounds to assume that the proficiency distributions vary over populations, and in these cases one may choose $\mu_0 = 0$ and $\sigma_0 = 1$ as the parameters of the prior distribution. The parameters κ_0 and v_0 can be seen as the number of *pseudo-observations* made. A noninformative prior distribution is obtained upon setting $\kappa_0 \rightarrow 0$, $v_0 \rightarrow -1$ and $\sigma_0^{-1} \rightarrow 0$.

So the procedure boils down to iteratively generating a number of sequences of parameter values using these six steps. Convergence can be evaluated by comparing the between- and within-sequence variance (see, for instance, Gelman et al., 1995). Starting points of the sequences can be provided by the Bayes modal estimates of Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996).

Incomplete Designs

In an incomplete design, only a subset of the total item pool is administered to each respondent. Let \mathbf{D}_j be the diagonal matrix of the elements $d_{1j}, \dots, d_{ij}, \dots, d_{nj}$, where $d_{ij} = 1$ if item j is administered to test taker i and $d_{ij} = 0$ otherwise. Furthermore, define $\mathbf{Z}_j = \mathbf{D}_j \mathbf{Z}_j^*$ and $\mathbf{X}_j = \mathbf{D}_j \mathbf{X}_j^*$. Then an incomplete design can be handled by changing

Step 1: if $d_{ij} = 1$ draw \mathbf{Z}_{ij} conditional on $\mathbf{W}_{ij}, \theta_i$ and ξ_j ;

Step 2: draw θ_i conditional on $d_{i1}, \dots, d_{ij}, \dots, d_{ik}$;

Step 3: draw ξ_j with \mathbf{Z}_j and \mathbf{X} replaced by \mathbf{Z}_j^* and \mathbf{X}_j^* , respectively;

Step 4: if $d_{ij} = 1$ draw \mathbf{W}_{ij} conditional on $\theta_i, \xi_j, \gamma_j$ and y_{ij} ;

Step 5: draw γ_j conditional on $d_{1j}, \dots, d_{ij}, \dots, d_{nj}$.

Modification Indices in a Bayesian Framework

In this article, three modification indices for the 2PL model in an MML framework will be generalized to the 3-PNO model in a fully Bayesian framework. These indices are the modification indices for the item characteristic curves and local independence (Glas, 1999) and the modification index for item parameter invariance, which is, for instance, applicable to evaluating differential item functioning and parameter drift (Glas, 1998b).

Consider the 3-PNO model 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. Then the null-model is estimated via MCMC as described above. When the Markov chain has converged, subsequent draws in the chain can be viewed as draws from the posterior distribution under the null-model. In every iteration of the MCMC procedure, parameters of the alternative model are drawn without consequences for the MCMC chain. So these draws do not influence the chain, and the Gibbs sampler remains restricted to the manifold of the posterior of the null-model without entering the support of the posterior under the alternative model. These draws will be called Bayesian modification (BM) indices. In this manner, many modification indices, for all items and several model violations, can be simultaneously sampled without mutual interaction. For every (set of) alternative model parameters, the ensemble of drawn Bayesian modification indices (BMD, Bayesian modification index distribution) gauges the peakedness of the alternative-model posterior outside the

manifold of the null-model. So this is analogous to the Lagrange multiplier modification indices in a frequentist approach, which gauge the change of the alternative model parameters if they should be set free, by measuring the peakedness of the likelihood function outside the manifold of the null-model. The BMD for a certain alternative model parameter cannot be viewed as some posterior distribution, but as an index of peakedness at the sampled places of the null-model posterior. Because the Markov chain stays in the manifold of the null-model, the draws of the alternative model parameters are weighted by their importance in terms of the null-model posterior density. Inferences with respect to the presence of a model violation can be made by computing the percentile of the value zero in the BMD. One could, for instance, decide that a certain violation required further investigation when the value zero was situated in the upper or lower 10% region of the generated BMD.

Application to the 3-PNO proceeds as follows: Above, η_{ij} was defined as the argument of the standard normal cumulative distribution function, that is, $\eta_{ij} = \alpha_j \theta_i - \beta_j$. Suppose that item j is the item of interest, and suppose the alternative model parameters are δ . Below, model violations will be modeled with the cumulative normal part of the 3-PNO, they could, in principle, also be modeled in conjunction with the guessing parameters, but this is beyond the scope of the present paper. A general model will be defined as

$$\eta_{ij} = \alpha_j \theta_i - \beta_j + \mathbf{x}_i^t (\zeta \theta_i - \delta), \quad (14)$$

where $\mathbf{x}_i^t (\zeta \theta_i - \delta)$ is the inner product of an observation \mathbf{x}_i and a vector that is a function of the alternative model parameters ζ and δ . It must be noted that this is a general formulation, in many applications the number of alternative parameters per item may be one only. After every execution of Step 3 of the above MCMC procedure, the parameters δ are sampled. From equation (14) it follows that under the alternative model, the definition of Z_{ij} can be extended to

$$Z_{ij} = \alpha_j \theta_i - \beta_j + \mathbf{x}_i^t (\zeta \theta_i - \delta) + \varepsilon_{ij}, \quad (15)$$

where ε_{ij} is a normally distributed error variable. Notice that equation (15) 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 ζ and δ as regression coefficients. So ζ and δ can be sampled in a manner analogous to Step 3 for the null-model. Notice that equation (15) also provides a nice interpretation of the procedure: the magnitude of the draws of ζ and δ depend on the extent to which the difference between Z_{ij} and $\alpha_j \theta_i - \beta_j$ is properly modeled under the null-model and the extent to which adding a predictor \mathbf{x}_i can improve the null-model.

Suppose that the null-model holds. Then the alternative model parameters in the vectors ζ and δ are all equal to zero, and, ideally, the number of positive values sampled should be approximately equal to the number of negative values. In other words, the proportion of positive values sampled, say the p-value, should be 50%. In practice, the p-value might be biased under the null-model. This will be returned to in the simulation studies reported below. If the model is violated, the p-value will become relatively high or low, and this can be regarded as an indication of the existence of a model violation. A significant modification index can be seen as caution index for further investigation. The mean and variance of the sampled values provide a further indication of the possible importance of the violation.

In the following section, this approach will be applied to derive three Bayesian modification indices for the 3-PNO model.

Bayesian Modification Indices for the 3-PNO Model

In this section, three Bayesian modification indices will be presented. They are focused at local independence, differential item function, and the item characteristic curves, respectively.

Local Independence

In the introduction section of this paper, in equation (2) an alternative to the 2PL model was presented where the axiom of local independence was relaxed by introducing a parameter δ_{jk} that modeled the dependence of the response to item k on the response to item j . This approach is brought within the framework of the previous section, by defining \mathbf{x}_i as a one-dimensional vector with an element y_{ij} and δ as a one-dimensional vector with an element δ_{jk} . The vector ζ is equal to zero. 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 . This assumption was relaxed in a model by Jannarone (1986). Translated to the present framework, the vector ζ then consists of a parameter ζ_{jk} modeling the relation of the dependence between the responses to θ .

Differential Item Functioning and Parameter Drift

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 & Wainer, 1993 or Camilli & Shepard, 1994) and various methods for detection of DIF have been proposed (Holland & Thayer, 1988; Hambleton & Rogers, 1989; Kelderman, 1989; Swaminathan & Rogers, 1990; Muraki & Bock, 1991).

Parameter drift has much in common with DIF. In both situations, one distinguishes between two or more groups of respondents. In DIF studies, one group serves as a reference group, and it is evaluated whether the response behavior of focal groups differs from the response behavior of the reference group. In studies of parameter drift, one may distinguish a calibration phase and a CAT phase and also here one evaluates whether response behavior differs. Therefore, it may come as no surprise that the statistical tools for the two kinds of studies are closely related.

Recently, Glas (1998) proposed a Lagrange multiplier modification index that is based on differences between the item parameters in a focal and reference population. The objective of this subsection is to generalize this index to a Bayesian framework. 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 modeled by adding parameters δ_j for modelling uniform DIF and δ_j and ζ_j for modelling non-uniform DIF, respectively.

The Form of the ICCs

For dichotomous items, Lord (1980, pp. 46–49) has pointed out that the expected number right score $\sum_i P(Y_{ij} = 1; \theta_i, \alpha_j, \beta_j, \gamma_j)$ and ability θ are the same things expressed on different scales of measurement. The important difference is that the measurement scale of the expected number right score depends on the test, while the measurement scale of θ is independent of the items in the test. Further, Grayson (1988) and Huynh (1994) have shown that, under the very mild conditions of unidimensionality, local independence, and nondecreasing ICCs, the number right score has a monotone likelihood ratio in θ . The idea of the modification index presented here is to partition the latent ability continuum into a number of segments, and to evaluate whether an item's ICC conforms the form predicted by the null-model in each of these segments. However, to be able to properly define the alternative model (see Glas, 1999), the observed total score scale rather than on the θ scale is partitioned. So let the item of interest be labeled j , while the other items are labeled $k = 1, 2, \dots, j-1, j+1, \dots, K$. Let $y_i^{(j)}$ be the response pattern without item j , and let $r(y_i^{(j)})$ be the unweighted sum score on this partial response pattern, that is,

$$r(y_i^{(j)}) = \sum_{k \neq j} y_{ik}. \quad (16)$$

The possible scores $r(y_i^{(j)})$ will be partitioned into S disjoint subsets; the index j signifies that this partition may be different for every item j . Consider the ordered boundary scores $r_0 < r_1 < r_2, \dots, < r_s < \dots, < r_S$, with $r_0 = 0$ and $r_S = K$. Further, define

$$w(s, \mathbf{y}_i^{(j)}) = \begin{cases} 1 & \text{if } r_{s-1} \leq r(\mathbf{y}_i^{(j)}) < r_s, \\ 0 & \text{otherwise,} \end{cases}$$

for $s = 1, \dots, S$. So $w(s, \mathbf{y}_i^{(j)})$ is an indicator function assuming a value equal to one if the unweighted sum score of response pattern $\mathbf{y}_i^{(j)}$ is in score range s . Because a partition of the score range also induces a partition of the sample of respondents, the term sub-sample will be used to signify groups of respondents

with a sum score in a certain subset of the score range. The choice of the number of subsets S and the choice of the boundary scores will be returned to below.

The essence of the approach is introducing an alternative model with discrimination parameters $\alpha_j + \zeta_{js}$ and difficulty parameters $\beta_i + \delta_{is}$. Consider a model where the probability of a correct score conditional on $w(s, \mathbf{y}_i^{(j)})$, is given by equation (14), with \mathbf{x}_i a $S - 1$ dimensional vector with elements

$$x_s = w(s, \mathbf{y}_i^{(j)}),$$

for $s = 1, \dots, S - 1$.

Under the null-model, ζ_{js} and δ_{is} will be equal to zero. In the alternative model, ζ_{js} and δ_{is} are free parameters, which gauge the deviation of the discrimination and difficulty parameters in the sub-groups from the values α_j and β_i . The model can be identified by the restrictions $\zeta_{js} = \delta_{is} = 0$. Under this parametrization, α_j and β_i are the discrimination and difficulty parameters in subgroup S and ζ_{js} and δ_{is} , $s = 1, \dots, S - 1$ are the deviations from this baseline in the other subgroups.

A Power Study

To investigate the power of the procedure proposed here, a number of simulation studies have been conducted. For the first simulations, the 3-PNO model holds and the item parameter values are given in the first columns of the first panel of Table 1. The test was administered to two groups. These groups had 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 1,000 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 score distributions per group, and the coefficient alpha per group.

TABLE 1
Simulation values and data summary—Number of observations = 2,000

Item	α_i	β_i	γ_i	p-value		Score		Frequency	
				$g = 1$	$g = 2$	$g = 1$	$g = 2$	$g = 1$	$g = 2$
1	0.40	-1.00	0.20	0.85	0.89	1	8	2	
2	0.60	-1.00	0.20	0.84	0.91	2	38	10	
3	0.80	-1.00	0.20	0.82	0.90	3	108	34	
4	0.40	0.00	0.20	0.60	0.66	4	191	103	
5	0.60	0.00	0.20	0.59	0.65	5	281	199	
6	0.80	0.00	0.20	0.60	0.69	6	358	309	
7	0.40	1.00	0.20	0.31	0.37	7	373	437	
8	0.60	1.00	0.20	0.33	0.40	8	316	435	
9	0.80	1.00	0.20	0.39	0.46	9	206	299	
10	0.60	0.00	0.20	0.60	0.70	10	96	143	
Group	μ_θ	σ_θ	Mean	S.D.	Alpha				
1	0.00	1.0	5.9	1.9	0.51				
2	0.50	0.8	6.7	1.7	0.39				

After computing the posterior distributions of the item and population parameters using MCMC, parameter estimates and confidence intervals can be 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 1,000 for the burn-in period and 3,000 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.

TABLE 2

Estimated parameter values and standard deviations—Number of observations = 2,000

Item	α_i	$sd(\alpha_i)$	β_i	$sd(\beta_i)$	γ_i	$sd(\gamma_i)$
1	0.41	0.07	-0.94	0.10	0.22	0.09
2	0.68	0.09	-0.98	0.10	0.23	0.09
3	0.93	0.13	-0.95	0.11	0.25	0.09
4	0.53	0.08	0.03	0.13	0.20	0.06
5	0.66	0.12	0.11	0.16	0.21	0.07
6	0.85	0.14	0.03	0.13	0.20	0.06
7	0.48	0.12	0.96	0.20	0.15	0.04
8	1.69	0.35	2.77	0.47	0.30	0.01
9	1.12	0.43	1.35	0.55	0.24	0.04
10	0.58	0.08	0.01	0.13	0.22	0.06
Group	μ_θ	$sd(\mu_\theta)$	σ_θ	$sd(\sigma_\theta)$		
1	0.00	—	1.00	—		
2	0.49	0.05	0.75	0.05		

Using the true item and population parameters of Table 1, a number of power studies were performed. These studies concerned the Type 1 error rate under the null-model, the detection rate of parameter drift and violation of local independence, say, the hit rate, and the Type 1 error rate under model violations, say, the false alarm-rate. Model violations were simulated by generating data with a non-zero value for a δ -parameter. Introducing non-zero ζ -parameters is beyond the scope of this study. For all studies reported below, 100 replications were made. It must be stressed that, given the number of replications, the reliability of the reported statistics is not very high, but the time consuming nature of the MCMC procedure impedes augmentation of the number of replications. In every replication, a BMD was considered significantly shifted from zero if the value $\delta = 0$ was positioned in the upper or lower 10% region of the BMD; that is, the generated distribution of δ -values. Power was defined as the percentage of significantly shifted BMDs over replications.

The studies on the Type 1 error rate under the null-model were conducted for sample sizes of 1,000, 2,000, and 4,000. Further, apart from a test length of 10 items, test lengths of 20 and 40 items were generated by duplicating the true parameter values of Table 1. In all combinations of sample size and test length, the Type 1 error rate power of all BM tests was under 1%. In a frequentist framework, it should be expected that the Type 1 error rate would be equal to the nominal significance probability, which in the present case would be 10%, but this reasoning is not quite appropriate in the present framework.

The second series of simulation studies entails the power of the test based on the BM for DIF and parameter drift. Data were generated using the true item and population parameters for Table 1, with the distinction that a DIF parameter $\delta = 0.50$ or $\delta = 1.00$ was added to item 10. The results for item 10 are shown in Table 3. As expected, the hit rate is an increasing function of both the sample and effect size. For the combination of the smallest sample and effect size, the power is definitely not very high. On the other hand, the number of false alarms for the non-affected 9 items was virtually zero.

TABLE 3

Power of the BM for parameter drift: Detection percentage

N	Effect Size	
	0.50	1.00
1,000	20	50
2,000	68	86
4,000	76	97

Also the BMDs for the forms of the ICCs and local independence were computed in this series of simulations. In the present version of the BM for the form of the ICC, within every one of the two groups of test takers, four score groups were formed. The score intervals were 0–2, 3–4, 5–6, and 7–9. Also this BM should be sensitive to DIF and parameter drift in item 10, and this is confirmed by the results shown in Table 4. Also in this case, the number of false alarms for the non-affected 9 items was virtually zero. Finally, the BMD for local independence was completely insensitive to DIF and parameter drift.

TABLE 4
*Power of the BM for ICC:
 Detection percentage for DIF
 items*

N	Effect Size	
	0.50	1.00
1,000	10	44
2,000	30	90
4,000	55	95

In the last series of simulations, the power of the BMD 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 $\delta = 0.50$ or $\delta = 1.00$ was added to model the dependency between item 9 and 10. The BMDs were computed for every consecutive pair of items. The results for the BMD of item 10 are shown in Table 5. The results shown in Table 5 are generally analogous to the results in the two previous tables, with the exception that the power for the combination of the smallest sample and effect size becomes negligible. The false alarm rate of the BMDs for the other items and the other two model violations considered here, were virtually zero.

TABLE 5
*Power of the BM for local
 independence: Detection
 percentage*

N	Effect Size	
	0.50	1.00
1,000	00	14
2,000	12	62
4,000	71	88

Discussion

In this paper, the use of a Bayesian framework to evaluate parameter drift was investigated. A number of Bayesian modification indices are presented that can be viewed as generalizations of the modification indices proposed by Glas and Verhelst (1995) and Glas (1998, 1999, 2001) for an MML framework. The strong point of the approach is that many model violations for all items can be assessed without complicated and time consuming computations. A weak point is that the power of the procedure is clearly inferior to the power of the analogous procedure in a marginal framework (see Glas, 1999). However, in real data situations, the power of the LM statistic is often quite high, so the lower power of the BM statistics may not be too much of a problem after all; these indices only serve the purpose of caution indices, and a significant result can be followed by a more detailed analysis with more traditional tools as posterior predictive checks and the Bayesian Information Criterion (BIC).

A point of further study is the generalization of the approach to more complex models. As mentioned above, the fully Bayesian approach applied to the 3-PNO does not really offer great advantages over the MML approach. The main advantage of the former approach is that it can be applied in complex IRT models where the latter approach breaks 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, & Wang, 1999) and models with multidimensional latent abilities (Béguin & Glas, 1998), and it is in the realm of these models that more research needs to be done.

References

- Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, 29, 813–828.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics*, 17, 251–269.
- Baker, F. B. (1998). An investigation of item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, 22, 153–169.

-
- Béguin, A. A., & Glas, C. A. W. (1998). MCMC estimation of multidimensional IRT models. *Statistical tests for person misfit in computerized adaptive testing* (Research Report 98-14). University of Twente, Enschede.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika*, *46*, 443–459.
- Box, G., & Tiao, G. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Fox, J. P., & Glas, C. A. W. (1998). *Multi-level IRT with measurement error in the predictor variables* (Research Report 98-16). University of Twente, Enschede.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.
- 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.
- Glas, C. A. W. (2001). *CUSUM statistics for large item banks: computation of standard errors* (Computerized Testing Report 98-11). Newtown, PA: Law School Admission Council.
- Glas, C. A. W. (2003). *Quality control of online calibration in computerized assessment* (Computerized Testing Report 97-15). Newtown, PA: Law School Admission Council.
- Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer and I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications*. (pp.69-96). New York: Springer.
- Grayson, D. A. (1988). Two-group classification in item response theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313–334.
- Hojtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs Sampler and posterior predictive checks. *Psychometrika*, *62*, 171–189.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, *59*, 77–79.
- Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika*, *51*, 357–373.
- Kelderman, H. (1984). Loglinear RM tests. *Psychometrika*, *49*, 223–245.
- Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*, 681–697.

-
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
- 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*. New York: Plenum Press.
- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177–195.
- Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data* [computer program]. Chicago: Scientific Software, Inc.
- Patz, R. J., & Junker, B. W. (1997). *Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses* (Technical Report No. 670). Pittsburgh: Carnegie Mellon University, Department of Statistics.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response theory models. *Journal of Educational and Behavioral Statistics*, 24, 146–178.
- 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.
- Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.
- Wainer, H., Bradlow, E. T., & Du, Z. (1999). Testlet response theory: An analogue for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computer adaptive testing: theory and practice*. Boston: Kluwer-Nijhoff.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Bilog MG: Multiple group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software.