ORIGINAL ARTICLE

An explanatory differential item functioning (DIF) model by the WinBUG 1.4

Saengla Chaimongkol[1], Fred W. Huffer[2] and Akihito Kamata[3]

Abstract

Chaimongko, S., Huffer, F. and Kamata, A.
An explanatory differential item functioning (DIF) model by the WinBUG 1.4
Songklanakarin J. Sci. Technol., 2007, 29(2) : 449- 458

This study proposed a multilevel logistic regression model to evaluate a source of DIF. The model accounts for the three levels nested structure of the data and combines results of logistic regression analyses to identify level-3 unit characteristic variables that explain a DIF variation. A simulation study is presented to assess the adequacy of the proposed models. The parameters of the proposed models were estimated by using a Bayesian approach implemented by the WinBUGS 1.4.

Key words: item response theory, multilevel logistic regression, bayesian

[1]Ph.D. (Statistics), Department of Mathematics and Statistics, Faculty of Science, Thammasat University, Khlong Lung, Phathum Thani, 12121 Thailand. [2]Ph.D. (Statistics), Florida State University, U.S.A., [3]Ph.D. (Educational Measurement and Statistics), Florida State University, U.S.A.
Corresponding e-mail: saengla@mathstat.sci.tu.ac.th, saenglastat@yahoo.com
Received, 31 April 2005 Accepted, 22 September 2005

บทคัดย่อ

แสงหล้า ชัยมงคล[1] เฟรด ดับเบิ้ลยู ฮัฟเฟอร์[2] และ อคีฮีโต๊ะ คามาต๊ะ[3]

ตัวแบบอธิบายการทำหน้าที่ต่างกันของข้อสอบโดยใช้วินบัก 1.4

ว. สงขลานครินทร์ วทท. 2550 29(2) : 449-458

การศึกษาครั้งนี้ได้เสนอตัวแบบพหุระดับการถดถอยโลจิสติกที่ใช้ในการตรวจสอบสาเหตุของการทำหน้าที่ต่างกันของข้อสอบ (DIF) โดยที่ตัวแบบที่เสนอนี้จะพิจารณาโครงสร้างของข้อมูลที่มีการซ้อนทับกัน 3 ระดับที่มีการรวมผลลัพธ์ที่ได้จากการวิเคราะห์การถดถอยโลจิสติก เพื่อที่จะระบุคุณลักษณะตัวแปรของข้อมูลระดับ 3 ที่สามารถใช้อธิบายสาเหตุการผันแปรของ DIF ได้ การศึกษาครั้งนี้จะใช้วิธีการจำลองข้อมูลในการตรวจสอบความถูกต้องและเหมาะสมของตัวแบบ โดยที่ค่าพารามิเตอร์ต่าง ๆในตัวแบบจะถูกประมาณโดยการใช้หลักการของเบย์ ที่ใช้โปรแกรมวินบักเวอร์ชั่น 1.4

[1]ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์ และเทคโนโลยีมหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12121 [2]ภาควิชาสถิติ มหาวิทยาลัยแห่งรัฐฟลอริด้า [3]ภาควิชาจิตวิทยาการศึกษาและการเรียนรู้ มหาวิทยาลัยแห่งรัฐฟลอริด้า เมืองทอลาฮาซี รัฐฟลอริด้า ประเทศสหรัฐอเมริกา

Differential item functioning (DIF) is present for a test item when respondents from two subpopulations with the same trait level have different probability of answering the item correctly. A consequence of having a DIF item is that the same true trait levels for examinees from different subpopulations could indicate different total test scores or trait level estimates. Currently, many statistical techniques have been proposed, based upon various theoretical backgrounds and practical purposes. A thorough review of DIF detection methods is given by Millsap and Everson (1993).

Once an item is identified as functioning differently from one subpopulation to another, understanding why the item is functioning differently between groups may be useful for many audiences. As one attempt, Gierl *et al*. (2003) studied gender DIF in mathematics by combining substantive and statistical analyses, as a two-stage process. Three difference statistical methods: SIBTEST, DIMTEST, and multiple linear regression, were used to test hypotheses about gender differences and to test whether content and cognitive differences were among items. Bolt (2000), for another example, found that multiple-choice items had more DIF characteristic than constructive-response items between males and females on SAT math pretest items. These results possibly suggest that knowing the source of DIF may be informative to minimize DIF items in future by many different means, including instruction, policy and test construction. These studies were based on multidimensional IRT based approaches.

Using another statistical approach, Swanson *et al*. (2002) proposed a two-level logistic regression model to evaluate sources of DIF. This approach explicitly accounts for the nested structure of the data and combines the results of logistic regression analyses across individual items to investigate the variation of DIF. Their level-1 models are a logistic regression model for DIF detection proposed by Swanminathan and Rogers (1990). In the level-2 models, the coefficients from level-1 models are treated as random variables and allow one to incorporate item characteristic variables to the models in order to explain the variation of DIF across items.

There is also a possibility that the magnitude of the DIF varies across group units, such as schools, and communities. Kamata and Binici (2003) first attempted to extend a two-level DIF model to three-level model using the hierarchical generalized linear model (HGLM) framework. The three-level model approach can be used to model variation of DIF across school as well as applied to identify the school characteristic variables that explain such variation. Their models were implemented by the HLM-5 software, which uses the penalized or predictive quasi-likelihood (PQL) method. They found that the variance estimates produced by the HLM-5 for the level 3 parameters are substantially negatively biased. This study extends their work by using a Bayesian approach to obtain more accurate parameter estimates. More specifically, this study will demonstrate a model in such a way that DIF of a particular item can be explained by some level-3 unit characteristic variables.

Methods

Model specification

To set the notation, let i denote the level-3 units (schools), j denote the level-2 units (students), and k denote the level-1 units (items). Assume that $i=1,2,...,M, j=1,2,...,n_i$ and $k=1,2,...,T$. Let y_{ijk} be the dichotomous response with code 1 if student j in school i responds to item k correctly and 0 otherwise.Let S_i denote the value of some characteristics for school i. For convenience, the values of S_i are often centered to have mean zero and may also be standardized to have unit variance. For all i , j and k , y_{ijk} are assumed to be independent Bernoulli random variables with the probability of correct response $P_{ijk} = P(y_{ijk} = 1)$. The explanatory DIF model can be written as:

$$y_{ijk} \sim \text{Bernoulli}(P_{ijk})$$

$$\text{logit}(P_{ijk}) = u3_i + u2_{ij} + \alpha 0 G_{ij} - \beta_k - \gamma 0_k G_{ij} + \delta 0_k S_i + \delta 1_k S_i G_{ij} + u4_{ik} G_{ij}, \quad (1)$$

where

$u3_i$ is the random effect for school i . It is assumed to be normally distributed with zero mean and constant variance (i.e.,$u3i \sim N(0,(\sigma 3)^2)$

$u2_{ij}$ is the ability of student in the school. It is a random effect and is assumed to be normally distributed with a non-zero mean and a group specific variance (i.e.,$u2_{ij} \sim N(\mu,\sigma^2_G)$).

$\alpha 0$ is the fixed effect of belonging to the focal group compared to the reference group, i.e., it is the mean difference between the focal and reference group.

G_{ij} is the group indicator that either indicates the reference group or the focal group. G_{ij} equals 0 if student j in school i belongs to the reference group and equals 1 if student j in school i belongs to the focal group.

β_k is a fixed effect representing the difficulty of item k for the reference group.

$\gamma 0_k$ represents the overall mean DIF for item k across schools.

$\delta 0_k$ is a fixed effect representing the main effect of the school characteristic for item k.

$\delta 1_k$ is a fixed effect representing an interaction between school characteristic and group for item k . In other words, $\delta 1_k$ indicates how much the DIF is different depending on the value of the school characteristics.

$u4_{ik}$ is a random increment to the DIF for item k in school i. It is assumed to be normally distributed with mean zero and item-specific variance (i.e.,$u4_{ijk} \sim N(0,(\sigma 4_k)^2)$ It is further assumed that the random effects $u3_i$, $u2_{ij}$, and are assumed to be mutually independent.

The values $\sigma 4_k$ provide a set of indices that describe how the DIF varies across schools. A large value of $\sigma 4_k$ indicates that, after controlling for school and student abilities, the DIF varies a great deal from school to school. On the other hand, a small or zero value of $\sigma 4_k$ indicates that the DIF varies little from school to school.

The model (1) is not identified because a constant can be added to all $u3_i$ and all β_k, but the logit of the model does not change. In addition, the parameter $\delta 0_k$ also causes an identifiability problem for the model since we can subtract a constant c from all the $\delta 0_k$ and add cS_i to $u3_i$ for all i without changing the logit. To identify the model (1), we adopt the approach of Bafumi *et al*. (2005) which suggested replacing the model parameters with new (adjusted) quantities that are well-identified but preserve the logit of the model. By following Bafumi's approach, adjust $\delta 0_k$ by subtracting its mean ($\bar{\delta}0$), and adding $\bar{\delta}0 S_i$ to all $u3_i$ to preserve the logit. Note that if S_i is centered to mean zero, then adding $\bar{\delta}0 S_i$ does not alter the mean of the $u3_i$. The other terms are adjusted as follows:

$$\begin{aligned} u3_i^{adj} &= u3_i - \bar{u}3 + \bar{\delta}0 S_i \\ u2_{ij}^{adj} &= u2_{ij} - \bar{\beta} + \bar{u}3 \\ \alpha 0^{adj} &= \alpha 0 - \bar{\gamma}0 \\ \gamma 0_k^{adj} &= \gamma 0_k - \bar{\gamma}0 \\ \delta 0_k^{adj} &= \delta 0_k - \bar{\delta}0. \end{aligned} \tag{2}$$

The original parameters are replaced by the corresponding adjusted parameters. The explanatory DIF model now can be defined as:

$$\text{logit}(P_{ijk}) = u3_i^{adj} + u2_{ij}^{adj} + \alpha 0^{adj} G_{ij} - \beta_k^{adj} - \gamma 0_k^{adj}{}_k G_{ij}$$
$$\delta 0_k^{adj} S_i + \delta 1_k S_i G_{ij} + u4_{ik} G_{ij,} \tag{3}$$

Estimation with WinBUGS

The explanatory DIF model can be easily implemented in WinBUGS, an available software for Bayesian analysis, using Gibbs sampling. The primary interest parameters are in the magnitude of $\delta 0_k$, $\bar{\delta}0$, and $\delta 1_k$. If the values of $\delta 1_k$ are substantial, the term of $\delta 1_k S_i$ explains a sizable part of the variability of DIF from school to school so that including this term in the model should cause a corresponding reduction in the value of $\sigma 4_k$.

The Bayesian approach treats all unknown parameters as random quantities with appropriate prior distributions. Estimation is based on the joint posterior distribution $P(\theta|y)$ where θ is the vector of unknown model parameters (i.e.,θ ($\{\alpha 0\}$, $\{\beta_k\}$,$\{\gamma 0_k\}$,$\{\delta 0_k\}$,$\{\delta 1_k\}$,μ,$\{\sigma 2_G\}$,$\sigma 3$,$\{\sigma 4_k\}$)) and y is the sample data. The posterior distribution of θ is obtained by Bayes' theorem as:

$$P(\theta|y) = \frac{P(y/\theta)P(\theta)}{\int P(\theta)P(y/\theta)d\theta}$$
$$\infty P(y|\theta)P(\theta).$$

where $P(\theta|y)$ is the likelihood and $P(\theta)$ is the prior distribution.

The posterior distribution is proportional to the likelihood multiplied by the prior distribution. Since the item responses given the school and student ability are assumed to be independent, the likelihood for the explanatory DIF model is given as:

$$P(y|\theta) = \int g(u4;0,\sigma 4_k) \int g(u3;0,\sigma 3)$$
$$\int g(u2;\mu,\sigma 2_G) \prod_{i,j,k} \int (y_{ijk} | u2_{ij}, u3_i u4_{ik})$$
$$du2_{ij} du3_i du4_{ik},$$

where

$$f(y_{ijk}|u2_{ij},u3_i,u4_{ik}) = \left(\frac{1}{1+e^{-n_{ijk}}}\right)^{y_{ijk}} \left(\frac{e^{-n_{ijk}}}{1+e^{n_{ijk}}}\right)^{1-y_{ijk}},$$

and $g(u4;0,\sigma 4_k)$, $g(u3;0,\sigma 3)$, and $g(u2;0,\sigma 2_G)$ are multivariate normal density of $u4 = \{u4_{ik}\}$, $u3 = \{u3_i\}$,and $u2 = \{u2_{ij}\}$,respectively, and η_{ijk} = logit(P_{ijk}) from equation (1).

The most difficult part of Bayesian inference is the complexity of the numerical evaluation of the posterior distribution because it involves integration and does not always produce a posterior and predictive distributions also involves high dimensional integrations. In recent years, these problems have been solved using Markov Chain Monte Carlo (MCMC) simulation methods, especially the Gibbs sampling algorithm implemented through the WinBUGS software, to simulate realizations from the joint posterior density. These samples are summarized to provide point and interval estimates for parameters in the model.

Choice of prior distributions and specification of initial values

Bayesian estimation of the model parameters requires the specification of a prior distribution for all the unknown parameters. In this study, a noninformative, but proper prior distribution is used. We assume the fixed effect ($\alpha 0$,$\{\beta_k\}$,$\{\gamma 0_k\}$, $\{\delta 0_k\}$,$\{\delta 1_k\}$), and μ of $u2$ are independent and normally distributed with mean zero and a huge variance10^4($N(0,10^4)$). For the variance parameters, we follow the recommendation of Gelman (2004) that suggests the use of a noninformative uniform prior density on standard deviation parameters unless a weakly informative prior is desired, in which case a half-t family such as a half-Cauchy prior distribution is recommended instead of a uniform prior. The half-t distribution can be defined as the ratio of the absolute value of a normal random variable centered at 0 and the square root of a gamma random variable. Further details on this distribution can be seen in Gelman (2004). For $\sigma 2_G$ and $\sigma 3$, a noninformative proper uniform prior density with a wide range (i.e., unif(0,1000)) are used. For the between-school standard deviation of DIF parameter for each item ($\sigma 4_k$), half-Cauchy prior distributions with scale parameter (ξ) of 25 as recommended by Gelman (2004) are used.

After setting the prior density for all unknown parameters, the model is completely specified in WinBUGS. Then WinBUGS loads the data and compiles the model. When the model compiles successfully, WinBUGS will load the initial values in the next step. For our model, we specify 0 as initial value for the fixed effects ($\alpha 0$, $\{\beta_k\}$,$\{\gamma 0_k\}$,$\{\delta 0_k\}$,$\{\delta 1_k\}$,and μ of $u2$, and 1 as the initial value for the standard deviation parameters of ($\sigma 2_G$,$\sigma 3$,$\sigma 4_k$). Using Gelman's (2004) approach to implementing the half-cauchy prior, the values of $\sigma 4_k$ are actually represented as

$$\sigma 4_k = \frac{|\xi_k|}{\sqrt{\tau_k}}$$

where $\xi_k \sim N(0,25)$ and $\xi_k \sim x_1^2$. We assign ξ_k and τ_k initial values of 1, so that $\sigma 4_k$ is also initially equal to 1. The initial values for the random effects ($u2$,$u3$,$u4$) are generated by WinBUGS itself.

When the initial values have been loaded or generated by WinBUGS successfully, WinBUGS now is ready to run the Gibbs sampling to obtain statistical inferences for the unknown parameters. In each situation we study, only one chain is run and the chain is run for 11,000 iterations with a burn-in of 1,000.

Simulation Study and Results

Simulation Design

To establish a realistic scenario, we generate 100 data sets from the model (1). There are 30 schools with a total of 1500 students. The number of students per school, n_i, varies and is chosen to be proportional to the number of students in each school. The n_i range from 9 to 103 students with an average of 50 students per school. We divide the students into reference and focal group equally with 750 students in each group. The test comprises 7 items with the difficulties β_k = 0.2,-0.2, 0.5, 0.8,-0.5, 0, -0.8, and DIF values $\gamma 0_k$ = -0.5, 0.0, -0.7, 0.0, 0.5, 0.0, 0.0. The mean difference between reference and focal group is $\alpha 0$ = 0.0, i.e., there is no mean difference between the two groups. We assume that only the first item is associated with the school characteristic S_i with main effect $\delta 0_1$=0.5 and the interaction with group $\delta 1_1$=0.5. For the other items,the values of $\delta 0_k$ and $\delta 1_k$ are assumed to be 0. The ability of students ($u2_{ij}$) is sampled from N(0,1). The ability of schools ($u3_i$) is sampled from N(0,.25). These parameter values are similar to what we observed in the data set of 2003 administration of a mathematics assessment for 4thgrade in a statewide testing program in the United States. The between-school DIF random effects $u4_{ik}$ are sampled from independent normal distributions with mean zero and standard deviations $\sigma 4_k$ = 0.5, 0.2, 0.5, 0.2, 0.2, 0.2, 0.2. The random effects $u2_{ij}$, $u3_i$ and $u4_{ik}$ are generated independently in each data set. It is assumed that $u2_{ij}$, $u3_i$ and $u4_{ik}$ are mutually independent. Splus is used to create the simulated data, and WinBUGS is used for the subsequent analysis.

Simulation Results

For each data set and analysis, our posterior inference is based on the output of a Gibbs sampler. We illustrate some typical Gibbs sample outputs. Sample history plot (trace plot), autocorrelation plot and posterior density plot are given for selected interest parameters for 10,000 iterations after eliminating the first 1,000 iterations. These plots are shown in Figures 1 to 3, respectively.

The trace plot is shown in Figure 1. Each parameter of interest becomes stationary by 1,000 iterations, indicating that convergence has been

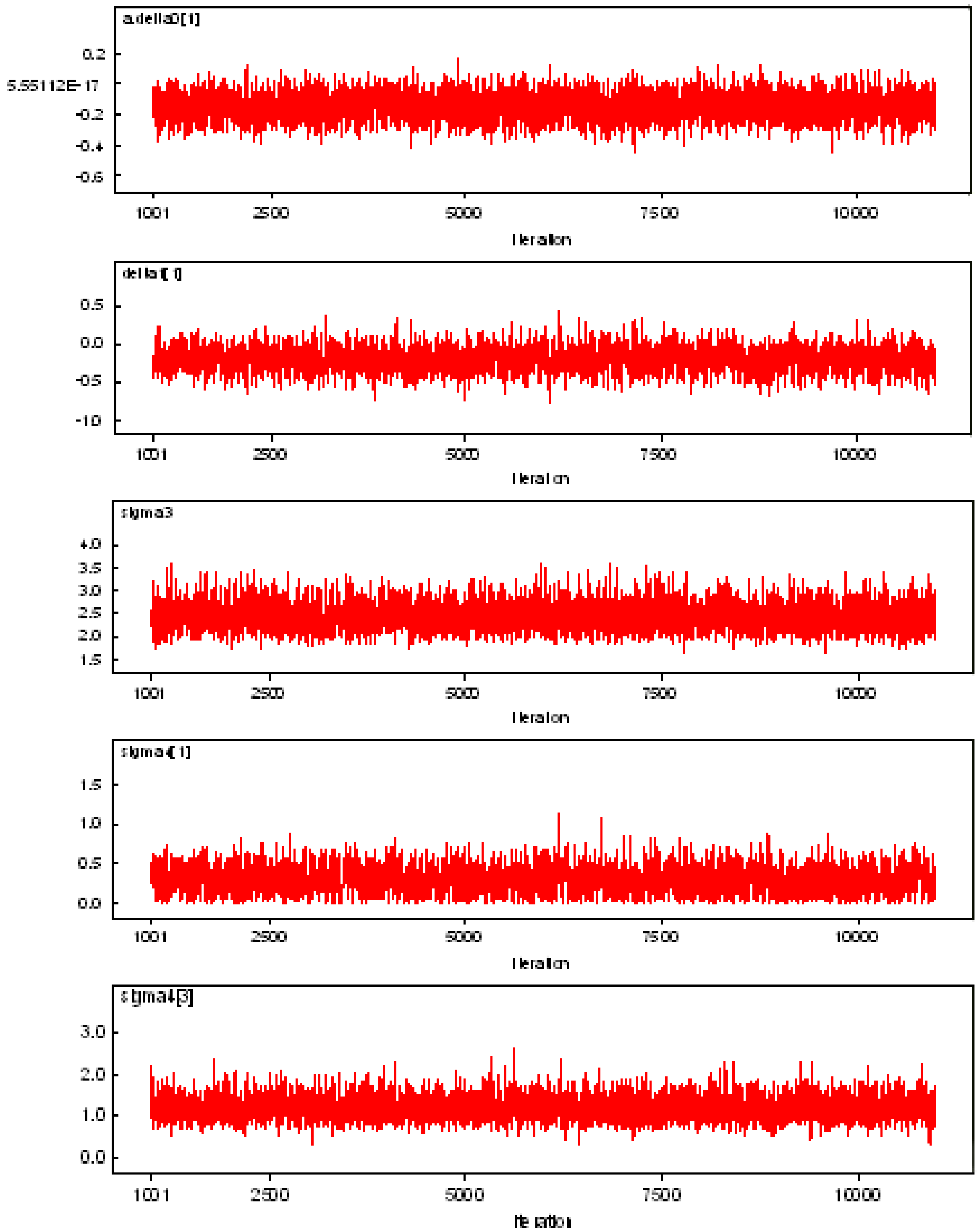

Figure 1. Gibbs sampling trace plots for some interesting parameters.

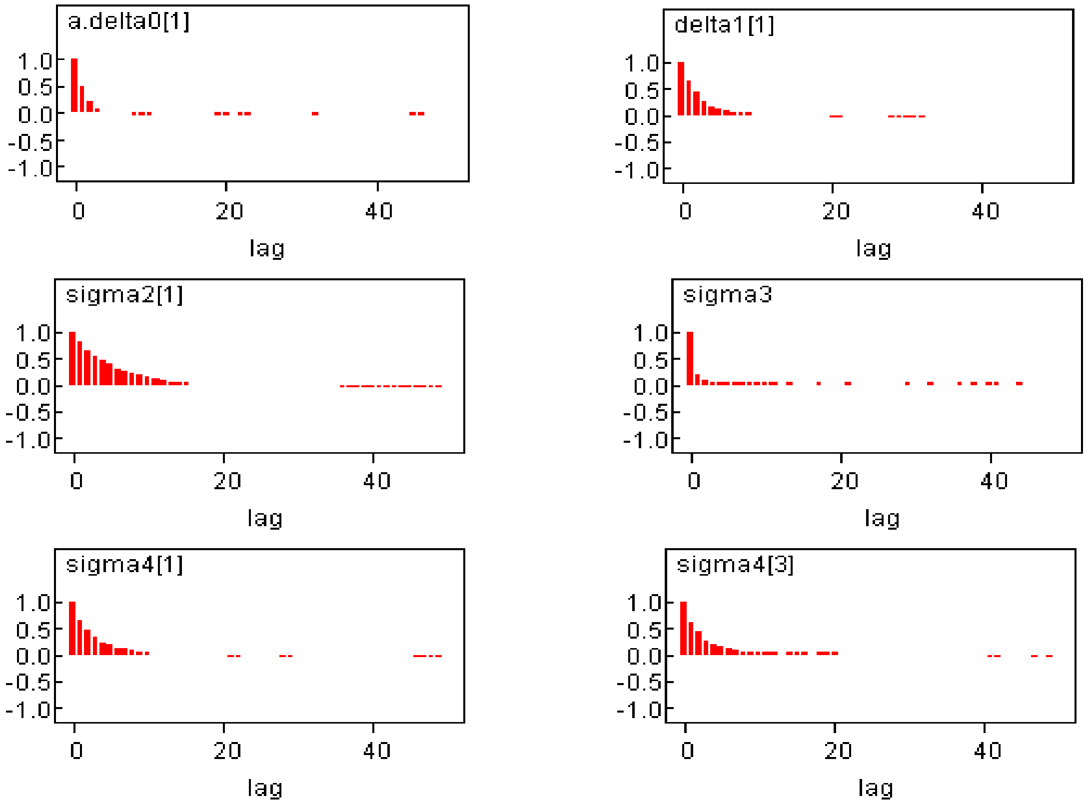

Figure 2. Autocorrelation plots for some interesting parameters

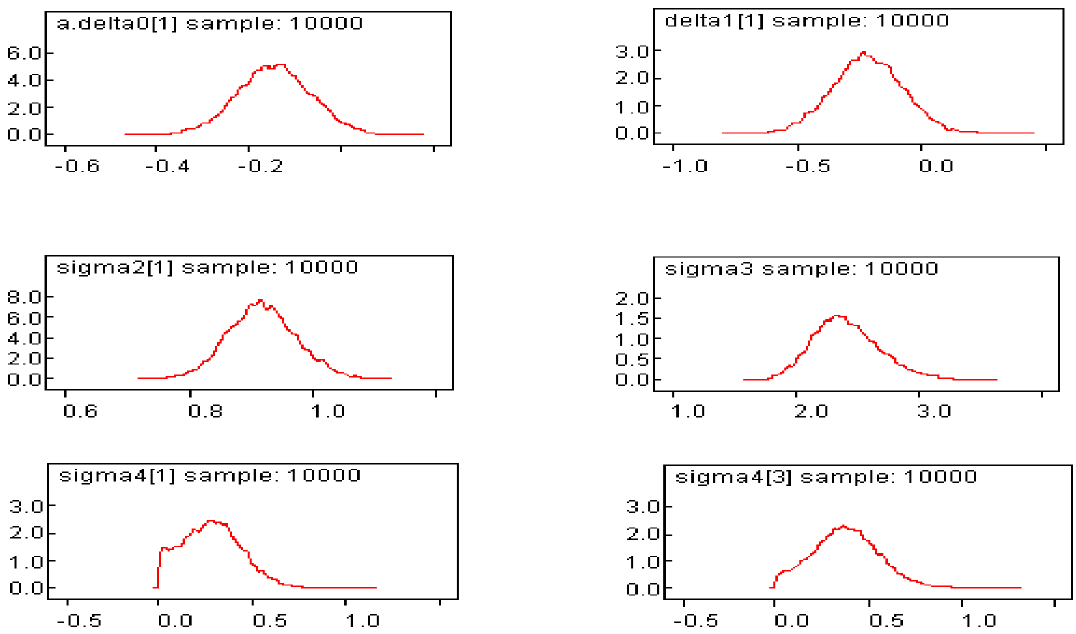

Figure 3. Posterior density plots for some interesting parameters

reached by 1,000 iterations. The autocorrelation plots (Figure 2) show that for all parameters, except the level-2 standard deviations (sigma2[1]), the autocorrelations decrease to near zero in fewer than l0 lags. The autocorrelations of the level-2 standard deviations approach to near zero by about lag 20. This indicates that the correlation between any two values separated by 10 or more iterations is close to zero, and these values can be treated as being roughly independent. These autocorrelation plots suggest that the chains are mixing well and quickly. In other words, the chains rapidly explore the entire posterior distribution.

The density plots for parameters (Figure 3) show unimodal distributions which are nearly symmetric, and look close to normal.

Table 1. Statistics of Gibbs sampling of interest fixed effect parameters in explanatory model.

Para-meter	Adj-True Value	Mean	SD	MCerror
$\delta 0_1^{adj}$	0.43	0.4356	0.0852	0.0014
$\delta 0_2^{adj}$	-0.07	-0.0702	0.0829	0.0015
$\delta 0_3^{adj}$	-0.07	-0.0800	0.0839	0.0015
$\delta 0_4^{adj}$	-0.07	-0.0788	0.0860	0.0015
$\delta 0_5^{adj}$	-0.07	-0.0654	0.0840	0.0015
$\delta 0_6^{adj}$	-0.07	-0.0661	0.0827	0.0015
$\delta 0_7^{adj}$	-0.07	-0.0751	0.0862	0.0015
$\delta 0$	0.07	0.0750	0.1146	0.0016
$\delta 1_1$	0.50	0.5222	0.1932	0.0048
$\delta 1_2$	0.00	-0.0085	0.1518	0.0034
$\delta 1_3$	0.00	0.0000	0.1784	0.0046
$\delta 1_4$	0.00	0.0085	0.1573	0.0035
$\delta 1_5$	0.00	0.0069	0.1534	0.0035
$\delta 1_6$	0.00	0.0336	0.1524	0.0035
$\delta 1_7$	0.00	0.0103	0.1587	0.0036

Each Gibbs sampler run produces 10,000 values for each parameter in the model. The sample mean and standard deviation (SD) of these 10,000 values estimate the posterior mean and standard deviation for that parameter, respectively. WinBugs also indicates the Monte Carlo error of our estimate of the posterior mean by estimating its standard deviation (referred to as the MC error below). The estimates of the posterior mean, standard deviation, and the MC error were computed for all 100 simulated data sets. These results are summarized in Tables 1 to 2. From these tables, we see that the mean over the 100 data sets is generally quite close to the true parameter value used in the simulations, except for the standard deviation parameters. Thus the point estimates given by the estimated posterior means are essentially unbiased. For the parameters $\sigma 4_k$, the mean of the point estimates is far from the true parameter value used in the simulations. Substantial bias exists when $\sigma 4_k$ is small (0.2 in our simulations). When the true value of the $\sigma 4_k$ is large enough (0.5 in our simulations), the bias is small (shown in Table 2).

When the model does not take into account the school variable, the summary statistics, including the mean, standard deviation (SD), and MC error of the posterior distribution are illustrated in Table 3. The mean point estimates of $\sigma 4_k$ from

Table 2. Statistics of Gibbs sampling of interest random effect parameters in explanatory model.

Para-meter	True Value	Mean	SD	MC error
$\sigma 2_1$	1.0	1.0117	0.0557	0.0016
$\sigma 2_2$	1.0	1.0245	0.0565	0.0016
$\sigma 3$	0.5	0.5265	0.0887	0.0020
$\sigma 4_1$	0.5	0.5352	0.1683	0.0020
$\sigma 4_2$	0.2	0.2397	0.1382	0.0030
$\sigma 4_3$	0.2	0.4975	0.1586	0.0034
$\sigma 4_4$	0.5	0.2465	0.1419	0.0030
$\sigma 4_5$	0.2	0.2554	0.1391	0.0030
$\sigma 4_6$	0.2	0.2512	0.1375	0.0030

Table 3. Statistics of Gibbs sampling of the standard deviation of DIF when the model does not take into account the school variable.

Para-meter	Mean	SD	MC error
$\sigma 4_1$	1.1509	0.2182	0.0047
$\sigma 4_2$	0.2448	0.1385	0.0030
$\sigma 4_3$	0.4943	0.1552	0.0032
$\sigma 4_4$	0.2447	0.1395	0.0029
$\sigma 4_5$	0.2566	0.1380	0.0030
$\sigma 4_6$	0.2444	0.1344	0.0028
$\sigma 4_7$	0.2560	0.1422	0.0030

this model are compared to the corresponding estimates from the explanatory model in Table 2. The results show that all of the $\sigma 4_k$ estimates of Table 3, except for $\sigma 4_1$ are close to the corresponding estimated by the $\sigma 4_k$ explanatory model.

Conclusion and suggestion

From the simulation results, the intuitive diagnostic tools, trace plots, autocorrelation plots, and posterior density plots, drawn from 10,000 updates after eliminating the first 1,000 suggest that convergence is achieved and the chains rapidly explore the entire posterior distribution. The point estimates given by the posterior means for fixed effect parameters, except the standard deviations, are essentially unbiased. The estimates of the standard deviation parameters are positively biased. This positive biased can be explained from the Bayesian point of view. Bayesian estimation combines the prior distribution with the likelihood to obtain the posterior distribution. When the data provide less information, the posterior distribution takes more heavily weights than the prior distribution resulting in shrinkage of the standard deviations toward the mean of the prior, which is a large positive value for a uniform distribution with a wide range from 0 to 1000 or a half-Cauchy distribution.

The skewness of the posterior density, when the true value of the standard deviation is small, suggests that the use of the posterior median or mode as a point estimate of standard deviation parameters may also reduce the relative bias for these parameters.

The mean of $\sigma 4_1$ is larger under the model that does not include the school variable than under the explanatory model. This makes sense since the explanatory model accounts for some of the school-to-school variability in the DIF by using the school characteristic S_i, whereas the model without school variable must include this variability in the random effects $u4_{ik}$ which inflates the value of $\sigma 4_1$.

This study is useful when the DIF exists at the individual level, and it interacts with the level-3 units. This is the case where the DIF is not consistent across level-3 units. If this is the case, knowledge that the DIF varies across level-3 units may be useful in revising the item since it alerts us to look for the differences between the level-3 units which may be cause the DIF.

References

Bafumi, J., Gelman, A., Park, D.K., & Kaplan N. (2005). Practical issues in implementing and understanding bayesian ideal point estimation. Political Analysis, 13, 171-187.

Bolt, D. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. JEM, 37, 307-327.

Gelman, A. (2004). Prior distributions for variance parameters in hierarchical models. Submitted. Downloadable from http://www.stat.columbia.edu/~gelman/research/published/tau9.pdf,[2004, April 10].

Gierl, M.J., Bisanz, J., Bisanz, G.L., & Boughton, K.A. (2003). Identifying content and cognitive skills that procedure gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. JEM, 4, 281-306.

Kamata, A., & Binici, S. (2003). Random-effect DIF analysis via hierarchical generalized linear models. Paper presented at the Annual International Meeting of the Psychometric Society, Sardinia, Italy.

Millsap, R.E., Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. APM, 17, 297- 334.

Swaminathan, H., & Rogers, H.J., (1990). Detecting differential item functioning using logistic regression procedures. JEM, 27, 361-370.

Swanson, D.B., Clauser, B.E., Case, S.M., Nungester, R.J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *JEBS*, 27, 53-75.