Choosing a Polytomous IRT Model using Bayesian Model Selection Methods

Hyun Jung Sung
Taehoon Kang
University of Wisconsin-Madison

March 29, 2006

Presented at the National Council on Measurement in Education annual meeting in San Francisco, April, 2006.

Running Head: IRT Model Selection Methods

Choosing a Polytomous IRT Model using Bayesian Model Selection Methods

Abstract

Model selection is the process by which a specific statistical model is chosen to represent the data. In order to get the benefits of item response theory (IRT), it is important to choose appropriate model which fit the data well. In this study, four model selection methods based on Bayesian estimation process will be compared in terms of their relative performances in choosing the best model to analyze Likert-type data. Among lots of polytomous IRT models already suggested, the rating scale model (RSM; Andrich, 1978), the partial credit model (PCM; Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the graded response model (GRM; Samejima, 1969) are used to compare the utility of the four model selection methods. Results indicate that model selection was dependent to some extent on the particular conditions simulated.

Index Terms: Item Response Theory, Model Selection, Polytomous IRT Model

1 Introduction

IRT is composed of a family of mathematical models designed to describe the relationship between examinee ability and performance on test items. Selection of an appropriate IRT model is critical if the benefits of IRT for applications such as test development, item banking, differential item functioning (DIF), computerized adaptive testing (CAT), and test equating are to be attained. Although there now exists an extensive IRT literature, relatively little has focused in methodology for determining the appropriateness of particular IRT models, and particular model comparison criteria. This has had the unfortunate consequence of many simply choosing a model with which they are familiar or for which software is available (Bolt, 2002; Embretson & Reise, 2000).

Because appropriate use of IRT models frequently depends heavily on model fit, the model selection process should be an important part in every application of IRT. If the wrong IRT model is selected for test data, the consequences can be severe in some cases. Yen (1981) explained the possible problems that could be caused by the use of an inappropriate model for dichotomous item response data. Perhaps most critically, the hallmark feature of IRT, parameter invariance, no longer applies (Shepard, Camilli & Williams, 1984; Bolt, 2002; Rupp & Zumbo, 2004).

Even beyond the practical implication of choosing an appropriate model, the model selection process can also help clarify the nature of the processes underlying test item responses. Many of the currently proposed IRT models differ according to how they characterize the nature of ability (e.g., dimensionality) and how they characterize the cognition mechanisms by which item scores are achieved (e.g., partial credit scoring versus graded response scoring). Because such insights are often a part of test validation, the methods studied in this dissertation may also assist IRT

researchers/practitioners in the process of determining whether their tests measure what they are designed to measure.

What is the best model? The *best* model can be defined in different ways depending on the goal of model selection. When the goal of model selection is only to find the model that provides the maximum fit to a given data set, a model with the smallest root mean squared deviation (RMSD) between the observed and the expected responses may be the best model. But, as Pitt, Kim, and Myung (2003) have noted, the goal of model selection can also be to identify the one model, from a set of competing models, that best captures the regularities or trends underlying the cognitive process of interest.

A more complicated model than appropriate violates the fundamental scientific principle of parsimony, which requires that one should choose the simplest of all the models that explain the data well. In the context of IRT, for example, if the only feature of interest were item difficulty, a model (such as the two parameter logistic model: 2PLM) which also adds an account of item discrimination might actually confuse understanding the item characteristic of interest. In brief, we want to choose the model that can explain all of the important features of the actual data without adding so much complexity that is unnecessary.

In this study, four model selection methods based on Bayesian estimation process are used in comparing IRT models. They are the deviance information criterion (DIC: Spiegelhalter, Best, & Carlin, 1998), the cross validation log-likelihoods (CVLLs) based on the concept of pseudo-Bayes Factor (PsBF: Geisser & Eddy, 1979; Gelfand & Dey, 1994; Bolt, Cohen & Wollack, 2001), and two information-theoretic methods which are Akaike's information criterion (AIC: Akaike, 1974) and Schwarz's Bayesian information criterion (BIC: Schwarz, 1978). All of them are known to be able to consider a model's complexity as well as its goodness-of-fit

(GOF: see Akaike, 1974; Forster, 1999; Kadane & Lazar, 2004; Massaro, Cohen, Campbell, & Rodriguez, 2001; Pitt, Kim, & Myung, 2003; Schwarz, 1978).

2 Polytomous IRT models

When items in a test are scored as one of more than two response categories like Likert-scale, polytomous IRT models are required. In this study, we deal with four commonly used polytomous IRT models: the rating scale model (RSM; Andrich, 1978), the partial credit model (PCM; Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the graded response model (GRM; Samejima, 1969). The first three models, the RSM, PCM, and GPCM, are hierarchically related, and represent an extension of "Binary Models" such as 2PLM in the Thissen and Steinberg (1986) taxonomy, referred to as "Divided-By-Total Models".

The most general of these three models is the GPCM. The probability that an examinee j scores in category x on item i is modeled by the GPCM as

$$P(X_{ij} = x|\theta_j, \alpha_i, \beta_i, \tau_{ki}) = \frac{\exp \sum_{k=0}^{x} \alpha_i[\theta_j - (\beta_i - \tau_{ki})]}{\sum_{y=0}^{m} \exp \sum_{k=0}^{y} \alpha_i[\theta_j - (\beta_i - \tau_{ki})]} \ , \tag{1}$$

where $j = 1, \ldots, N$, $i = 1, \ldots, T$, and $x = 0, \ldots, m$. In this model, α_i represents the discrimination of item i, β_i represents the difficulty of item i, and τ_k represents a location parameter for category k of item i. We set $\tau_{0i} = 0$ and $\exp \sum_{k=0}^{0} \alpha_i[\theta_j - (\beta_i - \tau_k)] = 1$ in Equation (1) for identification.

If the α_i is fixed at 1 across items, Equation (1) reduces to the PCM. In addition, if τ values are the same for each category, respectively, across items, Equation (1) further reduces to the RSM. Consequently, the RSM, PCM, and GPCM are nested models. Figure 1 shows example category response curves of a polytomous item with five categories (0, 1, 2, 3, and 4) under the GPCM. $\beta_i - \tau_1$ through $\beta_i - \tau_4$ indicate the spots in which the category response curves intersect on the latent-trait

scale.

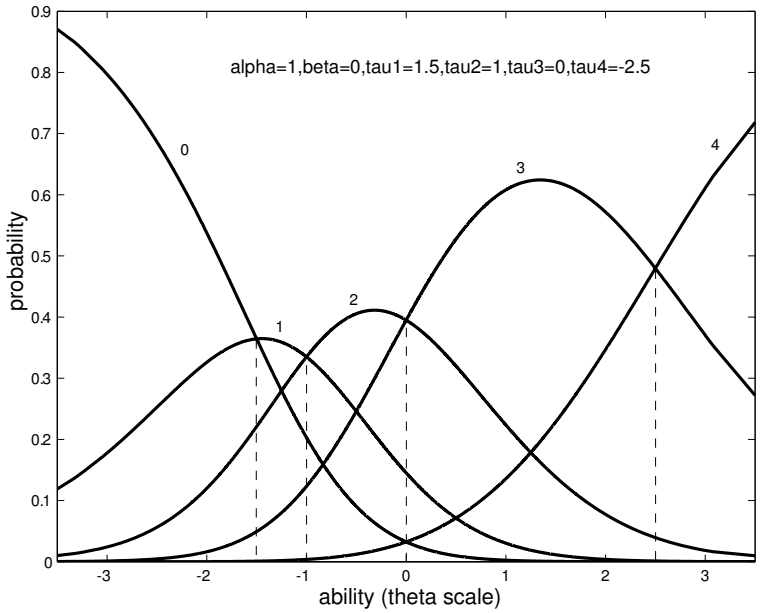

Figure 1: Category response curves for the example item under the GPCM: $\alpha = 1$, $\beta = 0$, $\tau 1 = 1.5$, $\tau 2 = 1$, $\tau 3 = 0$, and $\tau 4 = -2.5$

The GRM, however, is not a "Divided-By-Total" Model. Instead, the GRM is a representative model of another extension ("Difference Models") of Thissen and Steinberg's taxonomy. It can be viewed as a generalization of the 2PLM that uses the 2PL function to model boundary characteristic curves, namely curves that represent the probability of a response higher than a given category x. It is convenient in the model to convert the $x = 0, \ldots, m$ category scores into $x = 1, \ldots, m + 1$ categories. If we use P^*_{ijx} to denote the boundary probability for examinee j to have a category score larger than x on item i; then the boundary curve is given by

$$P^*_{ijx} = \frac{exp[\alpha_i(\theta_j - \beta_{xi})]}{1 + exp[\alpha_i(\theta_j - \beta_{xi})]} \ . \tag{2}$$

Figure 2 shows example boundary characteristic curves for a five-category item (1, 2, 3, 4, and 5) under the GRM. Note that an item with $m + 1$ categories results in

m boundary curves.

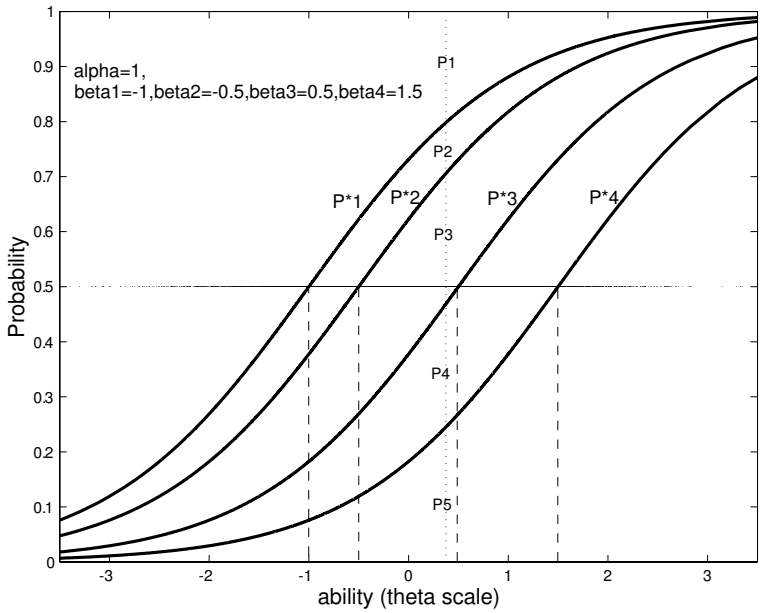

Figure 2: Boundary characteristic curves for the example item under the GRM

To determine the probability of a particular item score, the difference between adjacent categories is used. Thus, in the GRM, the probability that examinee j achieves category score x at item i is given by

$$P_{ijx} = P^*_{ij(x-1)} - P^*_{ijx} \tag{3}$$

where $x = 1, \ldots, m+1$, $P^*_{ij0} = 1$, and $P^*_{ij(m+1)} = 0$.

As an example, the values of P_{ij1} through P_{ij5} when the ability of examinee j is $\theta = 0.4$ is illustrated in Figure 2 as the length of vertical line divided by each boundary characteristic curves at the $\theta = 0.4$.

The GRM is distinguished from the GPCM and its nested models (the RSM, and PCM) by the fact that it requires a two-step process to compute the conditional probability for an examinee responding in a particular category. As Myung, Pitt, Zhang, and Balasubramanian (2001) explained, there are at least two independent

dimensions of model complexity: the number of free parameters of a model and its functional form (see $y = \theta x$ and $y = x^{\theta}$). Even though the GRM and GPCM need the same number of parameters for fitting each item, it is not easy to say that they have the same model complexity because the functions of the models are so different. Furthermore, the scoring process supposed by the GRM (grade response scoring) is conceptually different from that supposed by the PCM and GPCM (partial credit scoring). The former uses 2PLMs to compute boundary curves for each item, so each curve represents the probability of an examinee's raw item score (x) falling above a given category *threshold* as shown in Figure 2. (In fact, the β_{xi} in Equation (2) are often referred to as *threshold* parameters.) The order of category threshold should be kept within each item. In partial credit scoring, however, the focus is on the relative difficulty of each *step* needed to transition from one category to the next in an item. (Therefore, the $\beta_i - \tau_{ki}$ in Equation (1) are commonly referred to as *step* parameters.)

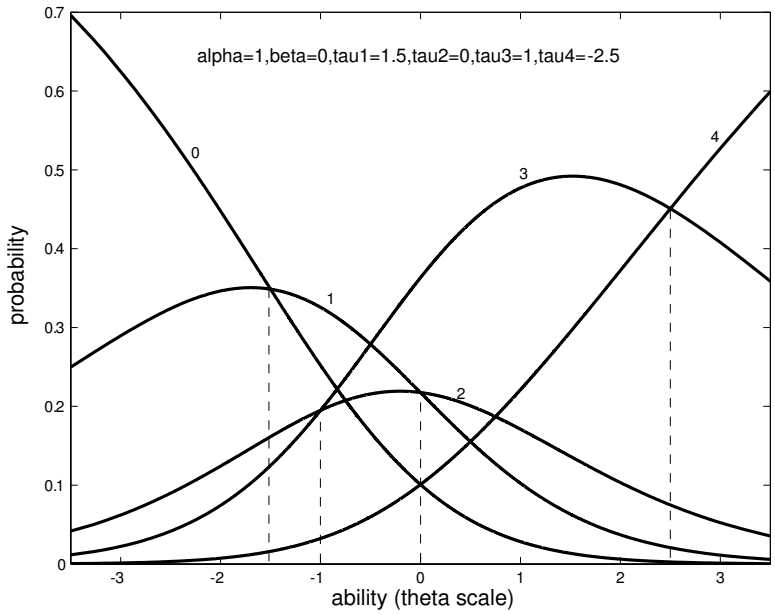

Figure 3: Category response curves for the example item under the GPCM: $\alpha = 1$, $\beta = 0$, $\tau 1 = 1.5$, $\tau 2 = 0$, $\tau 3 = 1$, and $\tau 4 = -2.5$

Within an item, some steps (category intersection) may be relatively easier or more difficult than others. So, the property of ordered location parameters is not indispensable. An example is illustrated in Figure 3 . There, the step from $x = 1$ to $x = 2$ (step parameter=0) is more difficult to achieve than that from $x = 2$ to $x = 3$ (step parameter=-1). The assumed scoring process for the RSM is differentiated from the PCM and GPCM in that the RSM model restricts such step processes to be same across all items in a test.

Bolt (2002) provided an illustration that demonstrated the necessity of selecting polytomous IRT models carefully in DIF analysis. To investigate the implication of model misspecification at DIF detection of polytomous response data, he conducted a simulation study to investigate the performances of the LR test under the GRM (referred to as GRM-LR test). Even though the GPCM and GRM appeared to provide similar GOF for a given data set, model misspecification had more serious implications for DIF analysis. When the best model for a given data set was the GPCM, but the GRM was used for model calibration and DIF detection, the GRM-LR test suffered from serious Type-I error inflation which would have been controlled if the correect model, GPCM, were used.

3 Model Selection Methods

It is often not clear to researchers and practitioners which one of polytomous IRT models provides the best description of the underlying item response process for a given set of data (Bolt, 2002). Therefore, techniques for distinguishing between these models are seemingly important, as is research on the benefits of choosing the best model and the problems with using a poor model.

Spiegelhalter et al. (2002) developed an index, DIC, to deal with Bayesian

posterior estimates of model parameters. DIC is composed of a Bayesian measure of fit or 'adequacy' called the posterior mean deviance $\bar{D}$ and a penalty for model complexity, p_D, the number of free parameters in the model.

$$DIC(Model) = \overline{D(\theta)} + p_D = D(\bar{\theta}) + 2 \times p_D, \tag{4}$$

where $\overline{D(\theta)}$, the posterior mean of the deviance, is a Bayesian measure of fit, $D(\bar{\theta})$ is the deviance of the posterior model (i.e., the deviance at the posterior estimates of the parameters of interest), and $p_D = \overline{D(\theta)} - D(\bar{\theta})$. The model with the smallest DIC is selected as the model that would best predict a replicate dataset of the same structure as that currently observed.

A common Bayesian approach to comparing two models, Model A and Model B, is to compute the ratio of the posterior odds of Model A to Model B divided by the prior odds of Model A to Model B. Bayes factor (BF) is the ratio of marginal likelihoods for the two models:

$$BF = \frac{posterior\ odds}{prior\ odds} = \frac{P(data|ModelA)}{P(data|ModelB)}. \tag{5}$$

A BF greater than 1.0 supports selection of Model A and a value less than 1.0 supports selection of Model B. It is known that Schwarz (1978) suggested BIC as an approximation to BF. According to Ghosh and Samanta (2001), Raftery (1995), and Western (1999), the difference between two BICs, $BIC_{ModelA} - BIC_{ModelB}$, is a fairly accurate approximation of $-2 \times log\ (BF)$, provided one of two models is a saturated model that fits the data perfectly.

The fact that the use of BF is only appropriate if it can be assumed that one of the models being compared is the true model (Smith, 1991) is a critical limitation on the common use for model selection. A less stringent assumption is that the two models are actually proxies for a true model. In this case, cross-validation log-

likelihoods (CVLL) can often be used to compute a PsBF to help determine which model to select (Spiegelhalter et al., 1996).

Below, it is explained how to calculate the CVLL in the IRT context. First, two samples are drawn, a calibration sample, $\mathbf{Y}_{cal}$ in which the examinees are randomly sampled from the whole data, and a cross-validation sample, $\mathbf{Y}_{cv}$, in which a second sample is randomly drawn from the remaining examinees. The calibration sample is used to update prior distributions of model parameters to posterior distributions. According to Bolt et al. (2003), the likelihood of the $\mathbf{Y}_{cv}$ for a model is then computed using the updated posterior distribution as a prior:

$$P(\mathbf{Y}_{cv}|Model) = \int P(\mathbf{Y}_{cv}|\theta, \mathbf{Y}_{cal}, Model) f_\theta(\theta|\mathbf{Y}_{cal}, Model) d\theta, \tag{6}$$

where $P(\mathbf{Y}_{cv}|\theta, \mathbf{Y}_{cal}, Model)$ represents the conditional likelihood, and $f_\theta(\theta|\mathbf{Y}_{cal}, Model)$ the conditional posterior distribution. An estimate of CVLL for a model is obtained as the logarithm of $P(\mathbf{Y}_{cv}|Model)$ in Equation (6).

The relationship between PsBF and CVLLs can be written as Equation (7), when Model A and Model B are being compared.

$$PsBF = exp(CVLL_A - CVLL_B). \tag{7}$$

The preferred model can naturally be determined through a direct comparison of individual CVLLs. When more than two models are compared together, the decision rule is that the model with the largest CVLL is the best (Spiegelhalter et al., 1996; Bolt et al., 2001). Estimates of CVLLs will be obtained using the MATLAB software. (An example of the MATLAB program used for this calculation is given in Appendix B).

Information-based indices are popular in many research areas because they strike a balance between the improvement in model fit by heavily parameterized model and

the elegance and predictability of a more parsimonious model (De Boeck, Wilson, & Acton, 2005). According to Sober (2002), Akaike's framework made us see the model selection problem in terms of the goal of predictive accuracy. Therefore, it came to be possible to pursue the *best* model under parsimony consideration based on observable evidences.

The AIC has two components representing GOF and complexity, respectively. The first component is the deviance (d) which is calculated with posterior means of item and ability parameters which are obtained by Gibbs sampler using the program WinBUGS 1.4 (Spigelhalter, Thomas, Best, & Lunn, 2003). And, the second component is $2 \times p$ where p is the number of estimated parameters, which can be interpreted as a penalty function for over-parameterization. This penalty is designed to correct for overfitting. The AIC is defined as:

$$AIC(Model) = d + 2p. \tag{8}$$

The model with the smallest AIC is the one to be selected. If a simple and a complex model fit a data set equally well, the simpler model will have the smaller AIC (Hitchcock & Sober, 2004). A criticism of the AIC is that it is not asymptotically consistent since sample size is not directly involved in its calculation (Ostini & Nering, 2005; Schwarz, 1978; Sclove, 1987). The AIC tends to prefer saturated models in very large samples (Janssen & De Boeck, 1999).

An alternate criterion similar to the AIC is the BIC. Schwarz (1978) developed the model selection measure, BIC, based on a Bayesian argument. The BIC achieves asymptotic consistency by penalizing over-parameterization with the use of a logarithmic function of the sample size. The BIC criterion is defined as

$$BIC(Model) = d + p \cdot (logN), \tag{9}$$

where N is the sample size. Whereas AIC multiplies p by a constant 2, p is multiplied

by a number proportional to N. Therefore, with the BIC, the penalty for increasing the number of parameters is more severe, particularly for data sets with large N. Not surprisingly, BIC tends to favor simper models relative to the AIC, when the sample size is large. As Lin & Dayton (1997) and Lubke & Muthén (2005) have noted, results from these two statistics do not always agree with each other because they have different penalties on the number of parameters.

4 Study 1: Comparison of Model Selection Indices on a Set of NAEP Mathematics Data

Methods for Study 1

There are two parts of this paper. Study 1 presents an example of the use of the four model selection indices obtained through Markov chain Monte Carlo (MCMC) algorithms. Study 2 presents a simulation study designed to explore the relative behavior of these indices on specific sets of data on the RSM, PCM, GPCM, and GRM.

Real Data. In Study 1, we present an example to illustrate the use of the four indices. Data for this study were taken from responses of Grade 8 students taking the 2000 State NAEP mathematics test. The 2000 State NAEP mathematics items were divided into 13 unique Blocks. Test booklets were developed for the 2000 State NAEP containing different combinations of three of the 13 Blocks. The design of the NAEP data collection ensured that each Block was administered to a representative sample of students within each jurisdiction (Allen et al., 1997). Students were allowed a total of 45 minutes for completion of all three Blocks.

Data from one of the 13 Blocks were used for this example. The Block had a total of 9 items, and 5 of those were scored polytomously as 0 (wrong), 1 (partially

correct), or 2 (correct). The GPCM was used to model the item response functions for this type of item (Allen et al., 1997). There were total 13,556 examinees who had took this Block.

Parameter Estimation. Bayesian parameter estimates were obtained using Gibbs sampling algorithms as implemented in the computer program WinBUGS. There is increasing attention to MCMC algorithms these days in IRT (see for example Baker, 1998; Bolt, Cohen, & Wollack, 2001; Kim, 2001; Patz & Junker, 1999a, 1999b, Wollack, Bolt, Cohen & Lee, 2002). In MCMC estimation, a Markov chain is simulated in which values representing parameters of the model are repeatedly sampled from their full conditional posterior distributions over a large number of iterations. The estimate is sampled from the posterior after each iteration. The value taken as the MCMC estimate is the mean over iterations sampled starting with the first iteration following burn-in. WinBUGS also provides an estimation of DIC for each set of items calibrated.

To derive the posterior distributions for each parameter, it is first necessary to specify their prior distributions. When there are 3 categories for each item, the following priors were used in this study. For the GPCM, $\theta_j \sim normal\ (0,\ 1), (j = 1, \ldots, N)$, $a_i\ \sim lognormal\ (0,\ 1), (i = 1, \ldots, T)$, $b_k \sim normal(0,\ 1), (i = 1, \ldots, T)$, $\tau_{1i} \sim normal\ (0,\ .1), (i = 1, \ldots, T)$, where N is the total number of examinees, T is the total number of items, a represents the discrimination parameter, b is the difficulty parameter, and τ_1 indicates the location of category 1 relative to the item's difficulty. For items with 3 categories (which are scored for NAEP as $x = 0,\ 1,\ 2$), the following constraints were used: $\sum_{k=0}^{m} \tau_{ki} = 0$, and $\tau_{2i} = -\tau_{1i}$ since $\tau_{0i} = 0$ in Equation (1). For the GRM, the following priors were used: $\theta_j \sim normal\ (0,\ 1), (j = 1, \ldots, N)$, $a_i \sim lognormal\ (0,\ 1), (i = 1, \ldots, T)$, $b_{1i} \sim normal\ (0,\ .1), (i = 1, \ldots, T)$,

$b_{2i} \sim normal\ (0,\ .1)I(b_{1i},), (i = 1, \ldots, T)$, where the notation $I(b_{1i},)$ indicates that b_{2i} is always sampled to be larger than b_{1i} (An example of the WinBUGS program used for calibration of the GRM is given in Appendix A).

Determination of a suitable burn-in was based on results from a chain run for a length of 11,000 iterations. The computer program WinBUGS (Spiegelhalter et al., 2003) provides several indices which can be used to determine an appropriate length for the burn-in. A previous study (Kang, Cohen & Sung, 2005) suggested that burn-in lengths of less than 100 iterations would be reasonable for any polytomous IRT model. A conservative estimate of 1,000 iterations for the burn-in was used in this study. For each chain, therefore, at least an additional 10,000 iterations was run subsequent to the burn-in iterations. Estimates of model parameters were based on the means of the sampled values from iterations following burn-in.

Results for Study 1

From the 2000 state NAEP mathematics test data, 3,000 examinees were randomly sampled for the calibration sample. Then, values for each of the four model selection indices were calculated. Also, to obtain the CVLL estimates, another 3,000 examinees were sampled from the same Block. Model selection results are reported in Table 1.

Table 1: Comparisons of model selection methods (2000 state NAEP math data: 5 polytomous items from Block 15)

	Model Selection Methods			
Model	DIC	CVLL	AIC	BIC
RSM	26005	-11950	24003	24039
PCM	23376	-10625	21424	21484
GPCM	22754	-10393	20894	20984
GRM	22774	-10292	20716	20806

The calibration sample consisted of 1,466 male and 1,534 female examinees. The minimum and maximum scores of the five polytomous item test were 0 and 10; the average score over all five items was 3.77 and the SD was 2.29.

With all the indices, the PCM and the RSM were ranked as the 3^{rd} and 4^{th}, respectively. There exist, however, some inconsistent results in terms of which is the best. The DIC for GPCM was the smallest, which means it chose the GPCM as the best model. The CVLL was the largest for GRM, and both the AIC and BIC had the smallest value for GRM. Therefore, these three indices selected the GRM as the best model. Given the lack of consistency, it is confusing to know which of these indices to apply in a practical testing situation.

5 Study 2: Simulation Study Comparing Model Selection Indices

In Study 2, we explore the behavior of these four indices further, using simulated data with known generating models and parameters. Here, data were generated with different IRT models under a variety of conditions. In this way, we hope to be able to better understand how the model selection indices might be used for model selection for conditions encountered in practical testing situations.

Methods for Study 2

Simulation Design. In Study 2, the design of the simulation study includes four polytomous IRT models described above (RSM, PCM, GPCM, and GRM), two test lengths ($n = 10$ or 20), two sample sizes ($N = 500$ or $1,000$), and two numbers of categories per item ($NC = 3$ or 5). The two test lengths are used to simulate tests having moderate and large numbers of polytomously scored items. The two sample sizes represent moderate and large samples. Discrimination parameters for the GPCM and GRM were randomly sampled from a lognormal(0, .5) distribution.

For five category items, item category parameters are randomly drawn from normal distributions with standard deviation of 1 and means of -1.5, -0.5, 0.5 and 1.5. After sampling, the difficulties were adjusted to meet the assumptions of each polytomous model. Threshold parameters for the boundary curves of the GRM must be ordered, so adjustments needed to be made when the randomly sampled thresholds did not result in ordered generating parameters. In such cases, the adjacent parameters were simply switched. For the GPCM, the mean of the item category generating parameters (b_{1i},..., b_{4i}) was used as the item difficulty parameter (b_i) and the difference between b_i and the b_{ki}s were taken as the step parameters, τ_{ki}s. θ values were randomly drawn from a normal $(0, 1)$ distribution.

For items with three categories, the location generating parameters were obtained as the mean of two adjacent generating parameters for the respective five category items. That is, the mean of b_{1i} and b_{2i} and the mean of b_{3i} and b_{4i} were taken as the new b_{1i} and b_{2i}, respectively, for items with three categories.

Table 2: Generating Item Parameters (NC=5)

	GRM					GPCM				
Item	a	b1	b2	b3	b4	a	b	τ_1	τ_2	τ_3
1	1.19	-1.59	-0.83	1.25	2.28	1.16	-0.42	2.56	-0.04	-1.67
2	0.96	-2.35	-0.29	0.60	1.84	0.51	-0.24	0.88	0.45	-1.67
3	1.52	-0.67	-0.06	1.28	2.39	1.43	0.61	3.05	-0.10	-0.95
4	2.48	-1.20	-0.04	1.22	2.42	2.25	-0.37	-0.41	1.88	0.00
5	0.58	-1.84	-1.13	-0.17	0.62	0.71	0.16	2.35	0.11	-0.67
6	1.13	-3.68	-2.23	-0.30	1.48	1.54	0.60	1.45	0.08	-0.26
7	1.63	-0.58	1.06	1.81	2.62	1.87	0.11	1.27	-0.24	0.50
8	0.82	-3.83	-0.98	0.49	1.12	0.45	-0.40	1.90	-0.60	-0.28
9	1.97	-3.51	-1.26	0.13	0.79	0.49	-0.38	3.17	-0.04	-2.08
10	1.21	-2.51	-1.65	0.72	1.62	1.33	0.15	1.59	-0.15	-0.34
11	1.10	-2.15	-1.40	0.59	1.48	0.82	-0.19	2.20	-0.38	-1.20
12	0.80	0.21	1.14	2.04	2.81	1.41	-0.03	0.73	0.60	-0.74
13	2.02	-3.07	-1.13	0.33	1.52	1.50	0.36	1.23	1.12	0.38
14	1.85	-0.64	0.22	1.00	1.83	1.43	0.35	0.03	1.02	0.28
15	1.48	-1.97	-0.03	0.96	2.41	1.91	-0.29	0.49	1.56	-1.36
16	1.40	-2.64	-1.30	-0.33	0.63	1.40	-0.34	1.68	0.27	-0.02
17	2.47	-2.09	-0.94	1.42	2.40	1.81	0.16	1.16	0.42	-1.24
18	0.93	-1.91	-0.79	0.44	1.26	0.55	-0.25	2.14	-0.18	-1.44
19	1.24	-1.61	-0.66	1.66	2.85	0.99	0.21	1.60	-0.86	0.41
20	1.65	-2.05	-0.16	0.67	1.96	0.92	0.19	1.62	0.92	-0.16
Mean	1.42	-1.98	-0.62	0.79	1.81	1.22	0.00	1.53	0.29	-0.63
SD	0.53	1.07	0.86	0.68	0.70	0.53	0.33	0.92	0.71	0.79

Table 3: Generating Item Parameters (NC=3)

	GRM			GPCM		
Item	a	b1	b2	a	b	τ_1
1	1.19	-1.21	1.77	1.16	-0.42	1.26
2	0.96	-1.32	1.22	0.51	-0.24	0.66
3	1.52	-0.36	1.84	1.43	0.61	1.47
4	2.48	-0.62	1.82	2.25	-0.37	0.74
5	0.58	-1.49	0.22	0.71	0.16	1.23
6	1.13	-2.96	0.59	1.54	0.60	0.76
7	1.63	0.24	2.21	1.87	0.11	0.52
8	0.82	-2.41	0.81	0.45	-0.40	0.65
9	1.97	-2.38	0.46	0.49	-0.38	1.57
10	1.21	-2.08	1.17	1.33	0.15	0.72
11	1.10	-1.78	1.04	0.82	-0.19	0.91
12	0.80	0.68	2.43	1.41	-0.03	0.67
13	2.02	-2.10	0.93	1.50	0.36	1.18
14	1.85	-0.21	1.42	1.43	0.35	0.52
15	1.48	-1.00	1.69	1.91	-0.29	1.03
16	1.40	-1.97	0.15	1.40	-0.34	0.97
17	2.47	-1.51	1.91	1.81	0.16	0.79
18	0.93	-1.35	0.85	0.55	-0.25	0.98
19	1.24	-1.14	2.25	0.99	0.21	0.37
20	1.65	-1.10	1.31	0.92	0.19	1.27
Mean	1.42	-1.30	1.30	1.22	0.00	0.91
SD	0.53	0.92	0.68	0.53	0.33	0.33

Tables 2 and 3 show the item parameters used for data generation. At the left side of the table are the generating parameters for the GRM and at the right side are the generating parameters for the GPCM. To generate a data set for the PCM, only the b and τ parameters from the right side of the table were used, and a parameters were fixed at 1. To generate a data set for the RSM, the τs of Item 1 were used for all items on the test. The first 10 item parameters were used for generating the 10-item tests, and all 20 items were used for generating the 20-item tests.

There were a total of 32 different conditions simulated in this study (4 generating models $\times$ 2 test lengths $\times$ 2 sample sizes $\times$ 2 category lengths). Ten replications will be generated for each condition. For the each generated data set, parameters for the same four polytomous models were estimated using MCMC techniques. To evaluate the performance of the four model selection indices, the indices were compared with respect to the proportions of times each index selected the correct model. A good

model selection index ought to be able to identify the generating model as the best model with a high percentage.

Results for Study 2

Recovery of Item Parameters. Since the model-selection indices in this study were calculated based on estimated model parameters, we first checked the quality of recovery of the item parameter estimation by MCMC. Parameter recovery was evaluated using product moment correlations (r) between the generating and the estimated parameters. The recovery results for all parameters in the four polytomous IRT models were very good ($r \geq .89$). The recovery results for the GPCM and GRM are reported in Table 4.

Table 4: Correlation Between Estimated and Generating Item Parameters

			GPCM by MCMC					GRM by MCMC				
test length	sample size	# of categ.	a	b	τ_1	τ_2	τ_3	a	b1	b2	b3	b4
n=10	500	NC=3	0.97	0.98	0.89			0.95	0.97	0.97		
		NC=5	0.98	0.99	0.98	0.95	0.97	0.97	0.95	0.99	0.99	0.96
	1000	NC=3	0.98	0.97	0.94			0.98	0.99	0.99		
		NC=5	0.99	0.99	0.99	0.97	0.98	0.98	0.98	0.99	0.99	0.98
n=20	500	NC=3	0.97	0.96	0.93			0.96	0.98	0.97		
		NC=5	0.98	0.98	0.97	0.97	0.97	0.98	0.98	0.99	0.98	0.97
	1000	NC=3	0.99	0.99	0.96			0.98	0.99	0.99		
		NC=5	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.99	0.99	0.98

Model Selection. The frequencies of model selections for the four different indices (DIC, CVLL, AIC and BIC) are shown in Figures 4, 5, and 6. In these plots, the main effects of three factors in Study 2 (test length, sample size, and number of categories) were illustrated, respectively.

In Figure 4, the model selection frequencies are plotted for different test lengths ($n = 10$, and $n = 20$). Because the frequencies were calculated marginally, the total 40 data sets (10 data sets $\times$ 2 sample sizes $\times$ 2 number of categories) were considered in each plot. When the true model was GPCM, PCM, or RSM, the

four indices performed well in selecting the correct model. A clear improvement in correct model selection was evident for the longer test ($n = 20$). Regardless of test length, when the true model was GRM, the GPCM tended to be selected as the best in roughly a third of the data sets.

Figure 5 shows the model selection frequencies plotted by sample size ($N = 500$ and $N = 1,000$). When the true model was the GRM, the performance of DIC appeared to be better for the larger sample size: when $N = 500$, DIC selected the GPCM as the better model about 2/3 of the time, and when $N = 1,000$, DIC performed better, selecting the correct model, GRM, with approximately 93% ($= 37/40$) accuracy. When the true model was one of GPCM, PCM, and RSM, the four indices performed well irrespective of sample size.

In Figure 6, the performance of the model selection indices looked sensitive to the number of categories. When a test has five-category items, the DIC, CVLL, AIC, and BIC selected the true GRM with 68%, 80%, 95%, and 95% accuracy, respectively, compared to each 55%, 70%, 63%, and 63% accuracy for three-category item tests. When the true model was the GPCM or RSM, all four indices worked almost perfectly in finding the correct model in the conditions with five-category items. The performance of CVLL showed a slight improvement in finding the true PCM as the number of categories was larger.

Table 5 shows the frequency that each index selected each model in each condition of Study 2. For example, for the 10 replications in the 20-item, $N = 1000$, and the number of categories=5 condition which were generated with the RSM (see the very bottom row of Table 5), the DIC index selected the GRM 0 times, the GPCM 0 times, the PCM 0 times and the RSM 10 times as the best model. In this condition, in fact, all four of the indices consistently selected the true model, RSM.

From Table 5, it was evident that in the conditions with large sample size

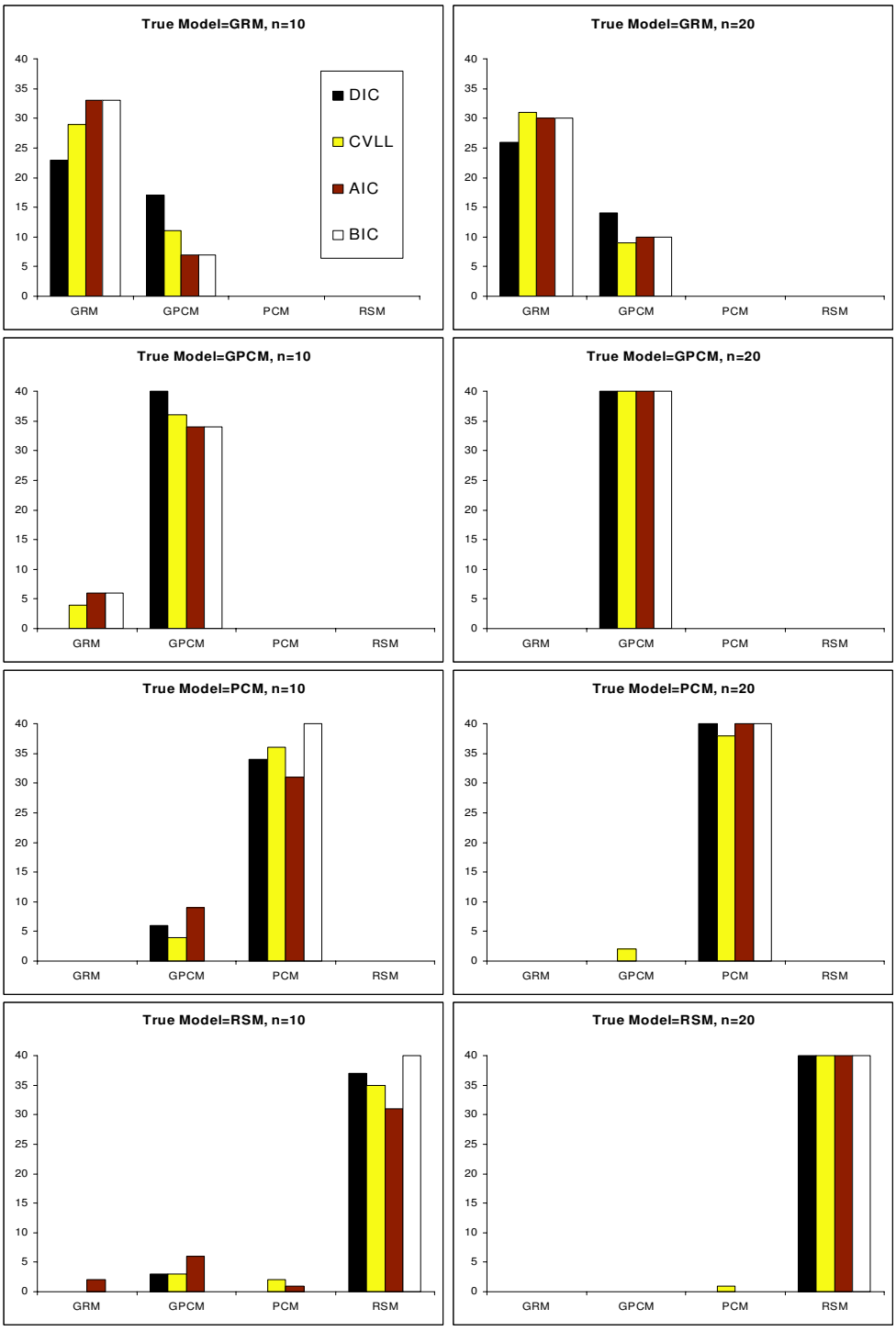

Figure 4: Model Selection Frequencies by Test Length

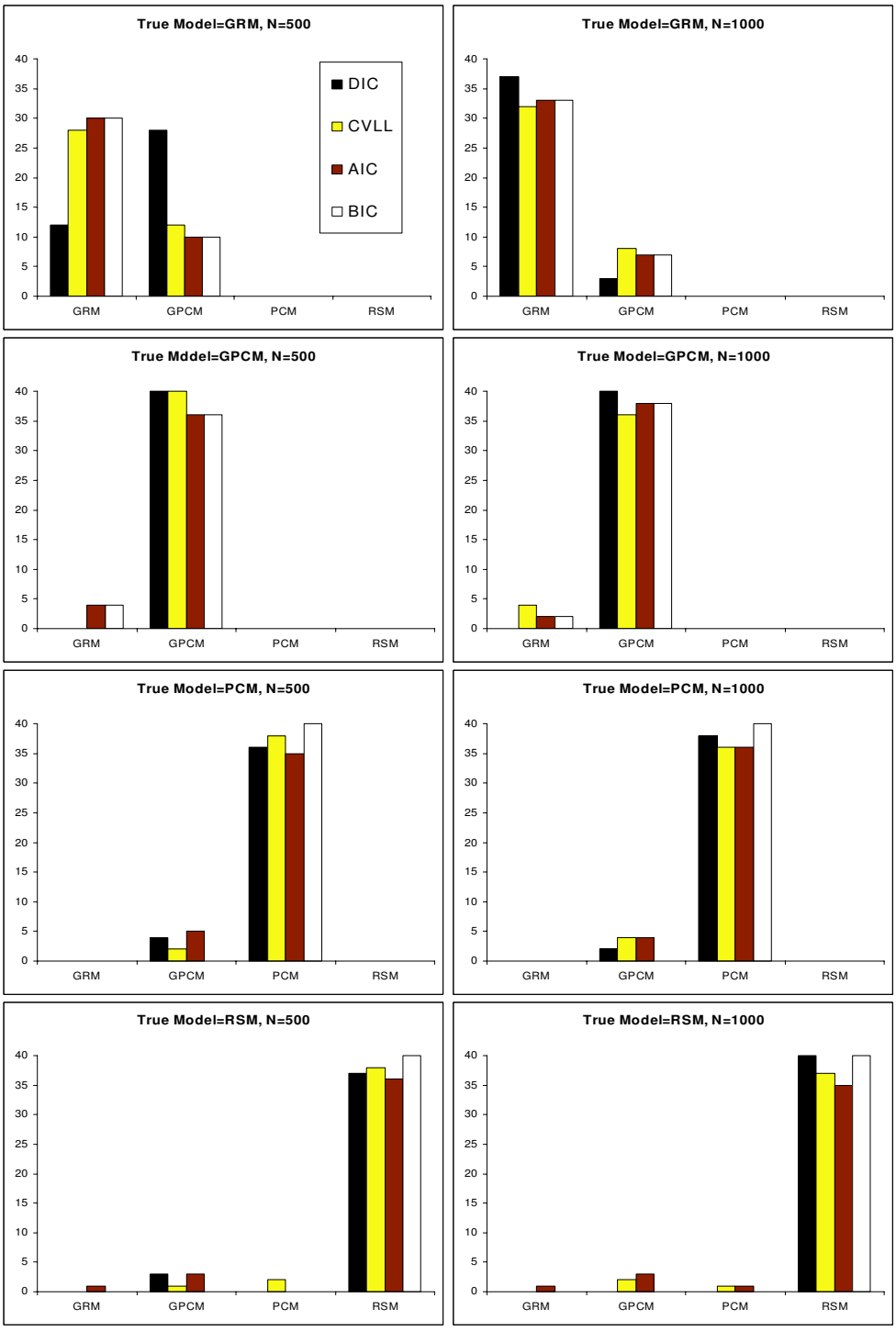

Figure 5: Model Selection Frequencies by Sample Size

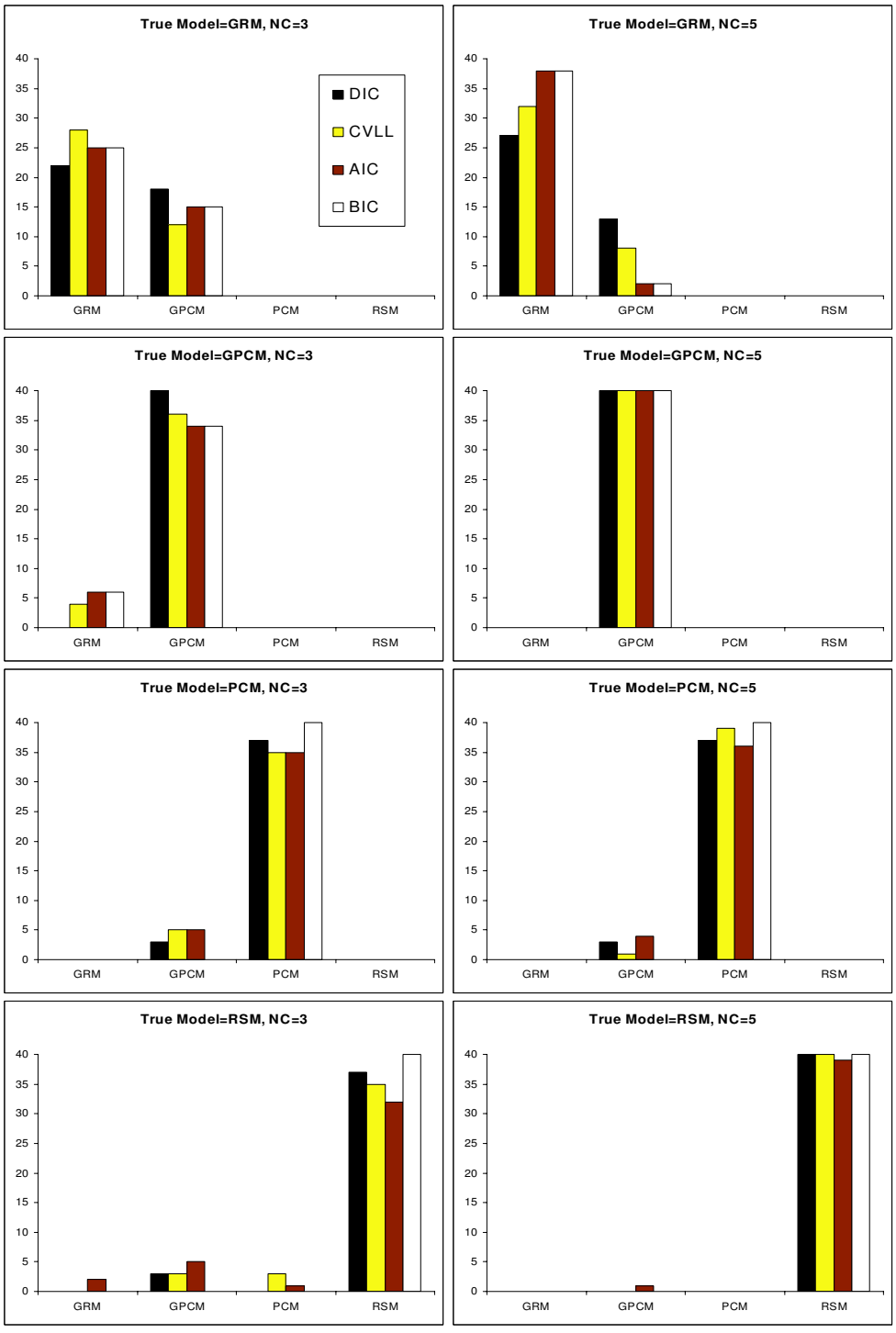

Figure 6: Model Selection Frequencies by Number of Categories

Table 5: Model Selection Frequencies

test leng.	samp. size	# of categ.	true* model	Selected by DIC GR	DIC GP	DIC P	DIC R	CVLL GR	CVLL GP	CVLL P	CVLL R	AIC GR	AIC GP	AIC P	AIC R	BIC GR	BIC GP	BIC P	BIC R
n=10	500	NC=3	GR	3	7	0	0	9	1	0	0	7	3	0	0	7	3	0	0
			GP	0	10	0	0	0	10	0	0	4	6	0	0	4	6	0	0
			P	0	1	9	0	0	1	9	0	0	2	8	0	0	0	10	0
			R	0	3	0	7	0	1	1	8	1	3	0	6	0	0	0	10
		NC=5	GR	2	8	0	0	2	8	0	0	8	2	0	0	8	2	0	0
			GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0
			P	0	3	7	0	0	1	9	0	0	3	7	0	0	0	10	0
			R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10
	1000	NC=3	GR	8	2	0	0	8	2	0	0	8	2	0	0	8	2	0	0
			GP	0	10	0	0	4	6	0	0	2	8	0	0	2	8	0	0
			P	0	2	8	0	0	2	8	0	0	3	7	0	0	0	10	0
			R	0	0	0	10	0	2	1	7	1	2	1	6	0	0	0	10
		NC=5	GR	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0
			GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0
			P	0	0	10	0	0	0	10	0	0	1	9	0	0	0	10	0
			R	0	0	0	10	0	0	0	10	0	1	0	9	0	0	0	10
n=20	500	NC=3	GR	2	8	0	0	7	3	0	0	5	5	0	0	5	5	0	0
			GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0
			P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0
			R	0	0	0	10	0	0	1	10	0	0	0	10	0	0	0	10
		NC=5	GR	5	5	0	0	10	0	0	0	10	0	0	0	10	0	0	0
			GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0
			P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0
			R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10
	1000	NC=3	GR	9	1	0	0	4	6	0	0	5	5	0	0	5	5	0	0
			GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0
			P	0	0	10	0	0	2	8	0	0	0	10	0	0	0	10	0
			R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10
		NC=5	GR	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0
			GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0
			P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0
			R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10

* GR=GRM, GP=GPCM, P=PCM, and R=RSM

(N=1,000) and 5 categories for each item, all four indices appeared to select the correct model almost perfectly. If the true model was one of the three nested models (RSM, PCM, and GPCM), all indices worked very well, especially when the test length is long (n=20). If the GRM was the correct model, however, CVLL, AIC, and BIC showed good performances only when the number of categories of items was 5. DIC, on the other hand, provided poor model choices for short test length (n=10) and small sample size (N=500).

Generally speaking, all four indices worked very well in almost all conditions at answering the correct model provided the data were generated with the RSM, PCM, and GPCM, but that indices often struggle to distinguish between GRM and GPCM when the data were generated with the GRM.

The simulated condition which was most similar to the real data from Study 1

was the n=10, N=1000, NC=3 condition. In Study 1, DIC had chosen the GPCM and the other indices had selected the GRM as the best model. If the true model for the NAEP data was the GRM which is of course unknowable, the simulation suggests that all four indices had a probability of 80% of selecting the correct model, GRM, according to Table 5. Also, if the true model for the data was the GPCM, the probabilities for choosing the generating model, GPCM, were 100%, 60%, 80%, and 80% for the DIC, CVLL, AIC, and BIC, respectively.

6 Discussion & Conclusions

There is no model that can perfectly describe a given set of data, because neither a theory nor a model can be a perfect mirror of reality (Wainer & Thissen, 1987). All we can do is faithful attempt to find the *best* model providing a sound connection between theoretical ideas and observed data (Navarro & Myung, 2005). Most studies of model-fit in IRT tend to focus on only GOF and item-model fit issues (e.g. Orlando & Thissen 2000, 2003; Glas & Suarez-Falcon, 2003). In terms of model selection considering the whole data at the same time, Forster (2004) named such approaches considering only model fitting as "naïve empiricism". And he warned that it would be problematic because it tends to indicate that the more complex model will be the better model, at least when the models are nested. This approach orients itself towards finding a model that fits the data perfectly. By doing so, the noise (idiosyncratic information) in the data will be fitted at the expense of the signal (structural information) behind the noise. Such "data dredging" may lead researchers to the discovery of spurious effects (Burnham & Anderson, 2002). This is why overfitting is undesirable. When we consider both GOF and model complexity, however, we can select a model with the best prediction accuracy. This

study was intended to help select a model among the four popular polytomous IRT models while keeping the principle of parsimony.

Deciding a model that generated a given data set is very difficult since the true model is not known for real data. In addition, data usually have random noise caused by sampling error, imprecise measurement instrument, or mistakes in data collection procedure. To make matters worse, it is always possible that more than one model could have generated the data sample. So then, what we have to do will be "to use all of the information available to make a best guess as to which model most likely generated the data (p. 351)" (Myung & Pitt, 2004). In this paper, it was assumed that there exists a generating or true model in the set of candidate models. However, what if there is no true model among them? When a researcher or a practitioner needs to find an appropriate model for his or her data set, the best he or she can do is to collect all available models with all the knowledge, experience, and help from other experts. Actually, this paper is intended to work in the situation where several competing models are already available through such process. From a divine point of view, however, what if there still could be a true model out of the candidate set? Then, maybe, it would be beyond the province of human beings with limited ability. What we can do is to try in earnest to select the best model among given models. Therefore, the criterion of best model should be how much the *predictive accuracy* can be obtained with a model, rather than whether or not a model is *true* (Forster, 1999; Hitchcock & Sober, 2004; Sober, 2002).

Maydeu-Olivares, Drasgow, and Mead (1994) used the ideal observer index (IOI) to compare the GPCM and GRM and concluded that either model would be equally appropriate in most practical applications. In terms of IRT model selection, two questions need to be answered related to their study. The first question is if it is really no use trying to choose one of the GPCM and GRM. Their conclusion is based

on the way they calculated the IOI. If the IOI following their way did not have enough power to distinguish two models, however, could we just believe their assertion that the two models would show same performances in most cases? Actually, Akkermans (1998) calculated the IOI in a different way and showed that it was possible to make the IOI much more powerful in finding the difference between the GPCM and GRM. It suggests the possibility that the two models may work differently. The other quesiton is if the IOI can be used for the purpose of model selection. As Ostini and Nering (2005) indicated, the computation of the IOI is not straightforward because it cab be estimated only with simulation data. Accordingly, although the IOI may be used to answer if any two statistical models perform differently, it cannot be a practical method for selecting an appropriate model for empirical data. So, the IOI was not considered in this paper.

As can be seen from the results of this study, inconsistencies and inaccuracies were found in model selection among the different indices in some of the simulated conditions. Some indices appeared to function better under some of the conditions than under others and for some models than for others. In general, it appears that for comparisons between the GRM and GPCM, the four indices were useful, when the true model was the GPCM. When the true model was the GRM, however, the performances of model selection indices were less accurate. Two interpretation will be possible for this phenomenon. One is the indices are just less powerful to find the true GRM. The other is that the GPCM is more flexible model than the GRM in spite of the use of the same number of parameters in modeling. In the study of Bolt (2002), when data sets were generated with the GPCM but the GRM-LR test was used for DIF detection, it was reported that there was a serious Type-I error inflation problem. But, Bolt did not deal with the opposite case where the the true model would be the GRM and the LR test could be done using GPCM. If the

second interpretation above is correct, we may be able to expect that less Type-I error inflation would happen in this approach. Further study seems warranted to make an investigation into this.

It is true that non-statistical issues in a model selection process are never trivial (as some models may be more appropriate for one type of psychological process than another or for one test purpose than another). Van der Ark (2001), actually, suggested to consider measurement properties of each competing polytomous IRT model in choice of an appropriate model. But, this study was intended to examine IRT model selection issue from a statistical perspective. All of the four model selection indices here were based on Bayesian item and ability calibrations. The results of studies such as this study may often be able to help inform a decision for selecting one model over another. However, additional study is necessary to enable a simulation study of this sort to provide practical guidance for researchers and practitioners. First of all, the simulating conditions in this study were too limited. Many more replications per condition (e.g. 50 or 100) need to be considered to produce more reliable and generalizable results. Also, there is no reason our interest is restricted within only the four models in this paper. Other IRT models for polytomous data such as sequential response model (Mellenbergh, 1995; Tutz, 1990) and unfolding model (Roberts & Laughlin, 1996) can be surely considered further on. In addition, more complicated IRT models dealing with multidimensionality may be investigated in terms of model selection.

References

Allen, N.L., Jenkins, F., Kulick, E., & Zelenak, C.A. (1997). *Technical report of the NAEP 1996 State Assessment Program in Mathematics.* Washington, D.C.: National Center for Education Statistics.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Anderson, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, (6), 716-723.

Ark, L.A. Van der. (2001). Relationship and Properties of Polytomous Item Response Theory Models. *Applied Psychological Measurement, 25*, (3), 273-282.

Baker, F.B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Baker, F. B. (1998). An Investigation of the Item Parameter Recovery Characteristics of a Gibbs Sampling Procedure. *Applied Psychological Measurement, 22*,(2), 153-169.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381-409.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo, *Applied Psychological Measurement, 27*, 395-414.

Box, G. E. P. (1976), Science and Education, *Journal of the American Statistical Association, 71*, 791-799.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic appraoch*, New York, NY, US: Springer.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 57*, 473-484.

De Boeck, P., Wilson, M., Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categoreis and dimensions, *Prychological Review, 112*, 129-158.

Dodd, B.G. (1984). Attitude scaling: A comparison of the grsded response and partial credit latent models (Doctorial dissertation, University of Texas at Austin). *Dissertation Abstracts International, 45*,2074A.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc., Publishers.

Forster, M. R. (1999). Model Selection in science: The problem of language variance. *The British Journal for the Philosophy of Science, 50*, 83-102.

Forster, M. R. (2004). Simplicity and unification in model selection. University of Wisconsin-Madison, WI. Available $http://philosophy.wisc.edu/forster/520/Chapter\%203.pdf$.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*, 153-160.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, B, 56*, 501-514.

Gill, F. (2002). *Bayesian Methods.* Boca Raton, FL: Chapman & Hall/CRC.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106.

Hitchcock, C., & Sober, E. (2004). Predictive versus accommodataion and the risk of overfitting. *British Society for the Philosophy of Science, 55*, 1-34.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory, *Multivariate Behavioral Research, 34*, 245-268.

Kadane, J. B., & Lazar, N. A. (2004). Methods and Criteria for Model Selection. *Journal of the American Statistical Association, 99*, 279-290.

Kang, T., Cohen, A.S. & Sung, H. J. (2005). IRT model selection methods for polytomous items. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Kang, T. & Cohen, A.S. (2004). IRT model selection methods for dichotomous items. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25*, 163-176.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22* (3), 249-264.

Lord, F. M. (1975). Relative efficiency of number-right and formula scores. *British Journal of Mathematical and Statistical Psychology, 28*, 46-50.

Loyd, B.H., & Hoover, H.D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

Lubke, G. H. & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review, 8*, 1-17.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Maydeu-Olivares, A., Drasgow, F. & D. Mead, A. (1994). Distinguishing Among Parametric Item Response Models for Polychotomous Ordered Data. *Applied Psychological Measurement, 18*, 245-256.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91-100.

Mislevy, D., J. & Bock, R., D. (1990). *BILOG: Item Analysis and Test Scoring with Binary Logistic Models.* Chicago, IL: Scientific Software. [Computer Program.]

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. & Bock, R. D. (1998). *PARSCALE (version 3.5): Parameter scaling of rating data.* Chicago, IL: Scientific Software, Inc.

Myung, I. J., Pitt, M. A. (2004). Model comparison methods. *Methods in Enzymology, 383), 351-366.*

Myung, I. J., Pitt, M. A., Zhang, S., & Balasubramanian, V. (2001). The use of MDL to select among computational models of cognition, in Leen, T. K., Dietterich, T. G., & Tresp, V. (Eds.), Advances in neural information processing system, Vol. 13 (pp. 38-44). MIT Press.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. it Applied Psychological Measurement, 24, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. it Applied Psychological Measurement, 27, 289-298.

Ostini, R. & Nering, M. L. (2005). Polytomous item response theory models. Thousand Oaks, CA:Sage.

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

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics, 24, 342-366.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. Psychonomic Bulletin & Review, 10, 29-44.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. Applied Psychological Measurement, 20, 231-255.

Rubin, D.B. (1984). Bayesinly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics, 12, 1151-1172.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. Educational and Psychological Measurement, 65, 588-599.

Sahu, S. K. (2002). Bayesian Estimation and model choice in item response models. Journal of Statistical Computation and Simulation, 72, 217-232.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, 17.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. Psychometrika, 52, 333-343.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Smith, A. F. M. (1991). Discussion of 'posterior Bayes factors' by M. Aitken. Journal of the Royal Statistical Society B, 53, 132-133.

Sober, E. (2002). Instrumentalism, Parsimony, and the Akaike Framework. Philosophy of Science, 69, 112-123.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. (1996). BUGS 0.5 Bayesian Inference Using Gibbs Sampling Manual (version ii). [Computer Program.]*

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical Report, MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology, 64, 583-616.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). WinBUGS 1.4 User Manual. [Computer Program.]*

The MATLAB 6.1 [Computer Software]. (2001). Natick, Massachusetts : The MathWorks, Inc.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item-response models. Psychometrika, 51, 567-577.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory. Chicago, IL: Scientific Software. [Computer Program.]

Tutz, G., (1990). Sequential item response models with an ordered response. British Journal of Statistical and Mathematical Psychology 43, 39-55.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. Applied Psychological Measurement, 25, 273-282.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. -S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and

Markov chain Monte Carlo estimation. Applied Psychological Measurement, 26, 339-352.

Western, B. (1999). Bayesian analysis for sociologists, Sociological Methods & Research, 28 (1), 7-34.

Yen, W. (1981). Using simulation results to choose a latent trait model. it Applied Psychological Measurement, 5, 245-262.

Appendix A: WinBUGS Code Used for GRM Calibration.

```
 grm.odc
-------------------------------------------------------------------------
# Graded Response Model

model {

    for (j in 1:N)  {
        for (i in 1:T)  {
        r[j,i]<-resp[j,i]
}}

# GRM
    for (j in 1:N)  {
        for (i in 1:T)  {
            for (k in 1: (mI[i]-1)) {
            p[j,i,k] <- 1 / (1+exp(-a[i]*(theta[j]-b[i,k])));
    }}}

    for (j in 1:N)  {
        for (i in 1:T)  {
        pcat[j,i,1] <- 1-p[j,i,1];
            for (k in 2: (mI[i]-1)) {
            pcat[j,i,k] <- p[j,i,k-1]-p[j,i,k];
            }
        pcat[j,i,mI[i]] <- p[j,i,(mI[i]-1)];
    }}

    for (j in 1:N)  {
        for (i in 1:T)  {
            for (k in 1:mI[i]) {
            pc[j,i,k] <- pcat[j,i,k]  / sum( pcat[j,i, 1:mI[i]] ) ;
            }
        r[j,i] ~ dcat(pc[j,i,1:mI[i]]);
        }
    theta[j] ~ dnorm(mu,1);
    }
mu ~ dnorm(0,1);

#  Priors
    for (i in 1:T) {
        a[i] ~ dlnorm(0, 1.);
        b[i,1] ~ dnorm(0,.1);
        for (k in 2: (mI[i]-1)) {
            b[i,k] ~ dnorm(0, .1) I(b[i,k-1], );
            }
       }
}
-------------------------------------------------------------------------
# The WinBUGS codes for calibration of other models can be
obtained from the authors
```

```
 cvlog.m
-------------------------------------------------------------------------
% information from posterior distributions of real data
N=3000; n=5;
load estrsm.txt;  % estimated item parameters by RSM
load estpcm.txt;  % estimated item parameters by PCM
load estgpcm.txt; % estimated item parameters by GPCM
load estgrm.txt;  % estimated item parameters by GRM
% CV dataset
load cv3000.dat; cvloglik=zeros(1,4);
% CV log-likelihood of RSM
a=ones(n,1);  b=estrsm(:,1); tau1=estrsm(:,2); tau2= -tau1;
cv=zeros(N,1); for j=1:N
    resp=zeros(1,n);
    resp=cv3000(j,:);
    ind_cv_gpcm
    cv(j)=cvj;
end cvloglik(1,1)=sum(cv);

% CV log-likelihood of PCM
a=ones(n,1);  b=estpcm(:,1); tau1=estpcm(:,2); tau2= -tau1;
cv=zeros(N,1); for j=1:N
    resp=zeros(1,n);
    resp=cv3000(j,:);
    ind_cv_gpcm
    cv(j)=cvj;
end cvloglik(1,2)=sum(cv);

% CV log-likelihood of GPCM
a=estgpcm(:,1);  b=estgpcm(:,2); tau1=estgpcm(:,3); tau2= -tau1;
cv=zeros(N,1); for j=1:N
    resp=zeros(1,n);
    resp=cv3000(j,:);
    ind_cv_gpcm
    cv(j)=cvj;
end cvloglik(1,3)=sum(cv);

% CV log-likelihood of GRM
a=estgrm(:,1);  b1=estgrm(:,2); b2=estgrm(:,3); cv=zeros(N,1); for
j=1:N
    resp=zeros(1,n);
    resp=cv3000(j,:);
    ind_cv_grm
    cv(j)=cvj;
end cvloglik(1,4)=sum(cv);
% PsBF
cvloglik
-------------------------------------------------------------------------

 ind_cv_gpcm.m
-------------------------------------------------------------------------
% 41 quadrature points between -4 to 4
k=-4:.2:4; K=length(k); prob=zeros(1,K); L=zeros(1,K);

% to calculate likelihood at each node
pofc=zeros(K,n,3); tt=zeros(K,n,3); denom=zeros(K,n); for t=1:K
        for i=1:n
        tt(t,i,1) = 1;
        tt(t,i,2) = exp(a(i)*(k(t)-b(i)-tau1(i)));
        tt(t,i,3) = exp(a(i)*(k(t)-b(i)-tau1(i) + k(t)-b(i)-tau2(i)));
        denom(t,i) = 1 + tt(t,i,2) + tt(t,i,3);
```

```
    end
end for t=1:K
    for i=1:n
        for w=1:3
            pofc(t,i,w)=tt(t,i,w)/denom(t,i);
        end
    end
end

for t=1:K
    lik=1;
    for i=1:n
        if resp(i)==1
            lik=lik*pofc(t,i,1);
        elseif resp(i)==2
            lik=lik*pofc(t,i,2);
        else
            lik=lik*pofc(t,i,3);
        end
    end
    L(t)=lik;
end

% to cmpute a posterior probability of ability
for t=1:K
    prob(t)=L(t)*normpdf(k(t),0,1);
end

prob=prob/sum(prob);

% to get CV log likelihood
cvj=0; for t=1:K
    cvj=cvj+prob(t)*log(L(t));
end
-------------------------------------------------------------------------

 ind_cv_grm.m
-------------------------------------------------------------------------
% 21 quadrature points between -4 to 4
k=-4:.4:4; K=length(k); prob=zeros(1,K); L=zeros(1,K);

% to calculate likelihood at each node
pofc=zeros(K,n,3); tt=zeros(K,n,2); for t=1:K
     for i=1:n
        tt(t,i,1) = 1/(1 + exp(-a(i)*(k(t) - b1(i) )));
        tt(t,i,2) = 1/(1 + exp(-a(i)*(k(t) - b2(i) )));
        pofc(t,i,1) = 1 - tt(t,i,1);
        pofc(t,i,2) = tt(t,i,1) - tt(t,i,2);
        pofc(t,i,3) = tt(t,i,2);
    end
end

for t=1:K
    lik=1;
    for i=1:n
        if resp(i)==1
            lik=lik*pofc(t,i,1);
        elseif resp(i)==2
            lik=lik*pofc(t,i,2);
        else
            lik=lik*pofc(t,i,3);
        end
    end
    L(t)=lik;
```

```
end

% to cmpute a posterior probability of ability
for t=1:K
    prob(t)=L(t)*normpdf(k(t),0,1);
end

prob=prob/sum(prob);

% to get CV log likelihood
cvj=0; for t=1:K
    cvj=cvj+prob(t)*log(L(t));
end
--------------------------------------------------------------------------
```