IRT Model Selection Methods for Polytomous Items

Taehoon Kang
University of Wisconsin-Madison

Allan S. Cohen
University of Georgia

Hyun Jung Sung
University of Wisconsin-Madison

March 11, 2005

Running Head: IRT Model Selection Methods

IRT Model Selection Methods for Polytomous Items

Abstract

A number of IRT models are available, each appropriate for a particular type of items or contexts. When such models are nested, the use of a likelihood ratio test may be appropriate for the purpose of model selection. When non-nested models or Bayesian estimation procedures are used, however, other item selection methods may be required. In this study, we compared five model selection indices, the likelihood ratio test, two information-based criteria, and two Bayesian methods, for use in model selection with nested and non-nested polytomous IRT models. In a simulation study, we compare the utility of these methods when models are nested and when models are non-nested. Results indicate that model selection was dependent to some extent on the particular conditions simulated.

Introduction

Item response theory (IRT) consists 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 based in part on model-data fit and is critical if the benefits of IRT are to be obtained. When IRT models are nested, it may be possible to select an appropriate model using a likelihood ratio (LR). The LR test statistic, G^2, is a chi-square based statistic and is calculated as $-2 \times \log$ of the likelihood for a given model. A difference between the G^2s for two models is itself distributed as a chi-square and so can be subjected to significance tests to determine which model is the better fit (Anderson, 1973; Baker, 1992; Bock & Aitkin, 1981).

When IRT models are not nested, an alternative approach is to investigate model selection. In such cases, it may be possible to use information-based statistics such as

Akaike's Information Criterion (AIC: Akaike, 1974) or Schwarz's Bayesian Information Criterion (BIC: Schwarz, 1978). Although significance tests are not possible with these statistics, they do provide estimates of the relative differences between solutions. These statistics are appropriate when maximum likelihood estimates of model parameters are obtained. As Lin & Dayton (1997), Lord (1975), and Sahu (2002) note, however, asymptotic estimates of item parameters may not always be available, in which case neither AIC nor BIC are appropriate. For such situations, Bayesian parameter estimation can sometimes be an effective alternative. Such estimates are obtained when using Markov chain Monte Carlo (MCMC) methods. Two model selection methods have been suggested when MCMC methods are used for estimation of IRT parameters: One is the pseudo-Bayes Factor (PsBF: Geisser & Eddy, 1979; Gelfand & Dey, 1994; Bolt, Cohen & Wollack, 2001), and the other is the Deviance Information Criterion (DIC: Spiegelhalter, D. J., Best, N. G., & Carlin, B. P., 1998). In this paper, we present some comparative evidence for efficacy of these different methods for selection of an appropriate IRT model.

As suggested above, comparisons among the different methods for model selection are complicated by the type of estimation, whether maximum likelihood or Bayesian, and by the relation among the models being considered, whether nested or not nested. If such comparisons could be made, they would provide useful information for making model selection decisions in practical testing situations. The LR test is appropriate only for comparisons among nested models. The AIC, BIC, PsBF, and DIC, however, may be used for comparisons of nested or non-nested models. This is important as non-nested models are frequently considered for modeling item response data. It would be of interest, therefore, to examine how the five model selection methods compare to one another. In this paper, we compare results for these five methods for use with data from polytomous item test.

To provide a basis for making comparisons among the five statistics, we will examine model selection using the following four 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 to each other. The probability for an examinee j to get a category score x at an item i is modeled by the GPCM as

$$P_{jix} = \frac{\exp \sum_{k=0}^{x} a_i[\theta_j - b_i + \tau_{ki}]}{\sum_{y=0}^{m} \exp \sum_{k=0}^{y} a_i[\theta_j - b_i + \tau_{ki}]} \ , \tag{1}$$

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

If the $a_i = 1$ across items, Equation (1) transforms to the PCM. If τ values are the same for each category, respectively, across items, Equation (1) further transforms to the RSM. So the first three models meet one of the conditions for the LR test, that is, they are nested. These models are not nested, however, with respect to the GRM. For model comparisons which include the GRM, therefore, the LR test would not be appropriate.

The GRM can be modeled using boundary characteristic curves to describe the probability of a response higher than category x. It is convenient to convert the $x = 0, \ldots, m$ category scores into $x = 1, \ldots, m + 1$ categories. P^*_{jix} is a boundary curve describing the probability for examinee j to have a category score larger than x on item i:

$$P^*_{jix} = \frac{exp[a_i(\theta_j - b_{xi})]}{1 + exp[a_i(\theta_j - b_{xi})]} \ . \tag{2}$$

Then, in GRM, the probability for an examinee j to achieve a category score x at item

i is

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

where $x = 1, \ldots, m+1$, $P^*_{ji0} = 1$, and $P^*_{ji(m+1)} = 0$.

Maximum likelihood algorithms are available for estimation of model parameters for all four models, meeting a condition for use of the LR, AIC and BIC statistics. If the GRM is to be compared to the RSM, PCM, and GPCM, however, the LR is no longer appropriate and other model selection indices need to be considered.

Parameters for these models can also be estimated using Bayesian algorithms and so meet a requirement for use of Bayesian model selection indices such as the PsBF and DIC. As suggested above, these comparisons are problematic, as maximum likelihood estimates of model parameters are required for the LR, AIC, and BIC statistics and Bayesian posterior estimates are needed for the PsBF or DIC statistics. One way to make comparisons among the five statistics would be to do so on common sets of data. Although different algorithms would be used for estimation of model parameters, at least they would be made on the same sets of data. Such comparisons would provide relative information about model selection among the different statistics. We illustrate the problem below on a set of State NAEP mathematics test data from 2000. In the sequel, we describe a series of simulations designed to provide relative information among the different statistics.

Model Selection Indices. The LR test tends to select a model with more parameters compared to models with fewer parameters. The availability of a significance test with the LR test, however, can be useful. The other four model selection statistics do not have associated significance tests. In addition, these other indices incorporate a kind of penalty on model complexity.

The AIC has two components. The first component, $-2 \times log\,(maximum \quad likelihood)$,

we refer to as G^2, the deviance. The smaller the deviance for a model, the better fitting the model. The other component, $2 \times p$, where p is the number of estimated parameters is intended as a penalty function for over-parameterization.

$$AIC(Model) = G^2 + 2p, \tag{4}$$

The model with the smallest is the one to be selected. A criticism of AIC is that it is not asymptotically consistent since sample size is not directly involved in its calculation. Although the AIC is useful, other indices such as the Bayes Factor (BF) also have been proposed.

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. BF is the ratio of marginal likelihoods for 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. One limitation in using BF is that it is only appropriate if it can be assumed that one of the models being compared is the true model (Smith, 1991). A less stringent assumption is that the two models are more appropriately regarded as proxies for a true model. In this case, cross-validation log-likelihoods can often be used to compute a PsBF to help determine which model to select (Spiegelhalter *et al.*, 1996).

Schwarz (1978) suggested BIC as an approximation to BF. According to Western (1999), the difference of two BICs, $BIC_{ModelA} - BIC_{ModelB}$, is a fairly accurate approximation of $-2 \times log\ (BF)$, where one of two models is a saturated model that fits the data perfectly. 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

$$BIC(Model) = G^2 + p\ (logN), \tag{6}$$

where N is the sample size. BIC tends to select models which are simpler than those selected by AIC. In other words, BIC gives a higher penalty to the number of parameters and thus tends to choose models with fewer parameters than the AIC. As Lin & Dayton (1997) have noted, results from these two statistics do not necessarily agree with each other.

Spiegelhalter, et al. (2002) proposed another index, the deviance information criterion (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.

$$DIC(Model) = \overline{D(\theta)} + p_D = D(\bar{\theta}) + 2 \times p_D, \tag{7}$$

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.

In Study 1, we present an example to illustrate the use of the five indices discussed above. Study 2 presents a simulation study designed to explore the relative behavior of these indices on specific sets of data for the four polytomous IRT models, the RSM, PCM, GPCM, and GRM.

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

Methods

Data. Data for this illustration will be 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 of items (Allen, Jenkins, Kulick, & Zelenak, 1997). 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 13 Blocks were used for this example. The block selected had a total of 9 items, 5 of which 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). Below, we compare the fit of the four models, the GPCM, the RSM, the PCM, and the GRM, on these data.

Parameter Estimation. Maximum likelihood estimates of item parameters were obtained using the computer program PARSCALE (Muraki & Bock, 1998). PARSCALE provides an estimate of $-2 \times log$ $(maximum\ likelihood)$ for each set of items calibrated.

Bayesian posterior parameter estimates were obtained using Gibbs sampling algorithms as implemented in the computer program WinBugs 1.4 (Spigelhalter, Thomas, Best, & Lunn, 2003). MCMC algorithms are receiving increasing attention in item response theory (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 1.4 also provides

an estimate of DIC for each set of items calibrated.

To derive the posterior distributions for each parameter, it was first necessary to specify their prior distributions. Items with 3 categories are modeled using the GPCM. The following priors were used for the GPCM: $\theta_j \sim normal(0,\ 1), (j = 1, \ldots, N)$, $a_i \sim (0,\ 1), (i = 1, \ldots, T)$, $b_k \sim normal(0,\ 1), (i = 1, \ldots, T)$, $\tau_{1i} \sim normal(0,\ .1), (i = 1, \ldots, T)$, where N is the total number of examinees, and 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 the 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). The priors for the RSM and PCM were subsets of these priors.

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}.

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. Preliminary results suggested that burn-in lengths of less than 100 iterations would be reasonable. Each of the chains actually converged relatively quickly to its stationary distribution, usually within the first 50 or so iterations. 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 were 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.

Cross Validation Log-Likelihood Estimates. Cross validation log-likelihoods

(CVLLs) were estimated for the PsBF method. Two samples were drawn, a calibration sample, $\mathbf{Y}_{cal}$ in which 3,000 examinees were randomly sampled from the examinees taking a given block, and a cross-validation sample, $\mathbf{Y}_{cv}$, in which a second sample of 3,000 were randomly drawn from the remaining examinees. Calculation of the CVLL proceeds first by using the calibration sample to update the prior distributions of model parameters to posterior distributions. Next, the likelihood of the $\mathbf{Y}_{cv}$ for a model can then be computed using the updated posterior distribution as a prior to (Bolt et al., 2003) :

$$P(\mathbf{Y}_{cv}|\mathbf{Model}) = \int P(\mathbf{Y}_{cv}|\theta_i, \mathbf{Y}_{cal}, Model) f_\theta(\theta_i|\mathbf{Y}_{cal}, Model) d\theta_i \ , \tag{8}$$

where $P(\mathbf{Y}_{cv}|\theta_i, \mathbf{Y}_{cal}, Model)$ represents the conditional likelihood, and $f_\theta(\theta_i|\mathbf{Y}_{cal}, Model)$ is the conditional posterior distribution.

Estimates of CVLLs were obtained using the MATLAB software. (An example of the MATLAB program used for this calculation is given in the Appendix.) To solve the integration in Equation (8), 41 Gaussian quadrature nodes were used in the MATLAB. As noted earlier, the BF index is defined relative to two different models. The PsBF estimated in this study used the two CVLLs. The best model is taken as the for which the CVLL is largest (Spiegelhalter et al., 1996; Bolt et al., 2001).

Results

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 indices, LR test, AIC, BIC, and DIC, were calculated. To obtain the CVLL estimates, another 3,000 examinees were sampled from the same block. Results are reported in Table 1.

Insert Table 1 About Here

The calibration sample consisted of 1,466 male and 1534 female examinees. The minimum and maximum scores over the five polytomous items on the test were 0 and 10; the average score over all five items was 3.77 and the SD was 2.29.

As shown in Table 1, the two Bayesian model selection methods identified the GRM as the best model for the data. The DIC for the GRM was smaller than for the other three models, and the CVLL for the GRM was the largest. The AIC and the BIC for the GPCM were smaller indicating the GPCM fit the data best. As was noted earlier, the LR test was appropriate for only the three nested, RSM, PCM, and GPCM. The LR test results suggested that the PCM fit better than RSM, and the GPCM fit better than the PCM.

Discussion for Study 1

The inconsistent results from Study 1 are somewhat consistent with previous research on these model selection indices. The LR test results suggested the more highly parameterized GPCM would be the best among the three hierarchically related models, GPCM, PCM, and RSM. The GRM, however, was identified as the best by the DIC and PsBF.

Study 2: Simulation Study Comparing Model Selection Indices

In the next study, we explore the behavior of these five indices further, using simulated data with known generating models and parameters. In this way, we hope to be able to better understand how these indices might be used for model selection for conditions encountered in practical testing situations, when nested or non-nested models are selected using model parameter estimates from either maximum likelihood or Bayesian algorithms.

Simulation Study Design.

In Study 2, we present a simulation study to compare the performances of the

five model selection methods. The design of the simulation study included two test lengths, 10 and 20 items, two sample sizes, 500 and 1,000 examinees, two numbers of categories per item, 3 and 5, and two distributions of ability, $N(-1, 1)$ and $N(0, 1)$. The two test lengths were used to simulate tests having small and large numbers of polytomously scored items. The two sample sizes were used to simulate small and large samples. Discrimination parameters for the GPCM and GRM were randomly sampled from a log-normal distribution, $ln(0, .5)$. Item category difficulty parameters were randomly drawn from normal distributions: $N(-1.5, 1)$, $N(-0.5, 1)$, $N(0.5, 1)$, and $N(1.5, 1)$. These distributions were used for five category items. After sampling, the difficulties were adjusted to fit them for to meet the assumptions of each polytomous model. Location parameters of the boundary curves in GRM must be ordered, so adjustments needed to be made when the randomly sampled ones did not result in ordered generating parameters. In such cases, the adjacent parameters were simply switched. And, for GPCM as Equation (1), 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 b_{ki} was taken as the step parameter, τ_{ki}. An additional adjustment was needed to ensure that the average of item difficulties for a simulated test was zero. We were thus able to simulate samples of examinees whose ability distributions did or did not match the difficulty of the test. For the items with three categories, the location generating parameters were obtained as the mean of two adjacent generating parameters from the generating parameters sampled for the respective five category items. In other words, the mean of b_{1i} and b_{2i} and the mean of b_{3i} and b_{4i} became the new b_{1i} and b_{2i}, respectively, for items with three categories.

Table 2 shows 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. 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.

Insert Table 2-1 and 2-2 About Here

There were a total of 64 different conditions simulated in this study (2 test length $\times$ 2 sample sizes $\times$ 2 category length $\times$ 2 ability distributions $\times$ 4 true models). Ten replications were generated for each condition. Item parameter estimation was as described for Study 1.

Simulation Study Results

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 estimates. Parameter recovery was evaluated using both the root mean square error (RMSE) and product moment correlations (σ) between the generating and the estimated parameters. Before calculating RMSEs, the estimated parameters were linked to the generating parameter scale using the mean-mean procedure (Loyd & Hoover, 1980).

Insert Table 3-1, 3-2, 3-3, and 3-4 About Here

The results in Table 3 are in general agreement with recovery results reported in the literature. In Table 3-1, the recovery results for the GPCM are presented. Most σs were larger than .95 and the RMSEs were around .1 for both MMLE and MCMC.

Two very large RMSEs (1.96 and 2.19) and two very small σs (0.26 and 0.46) were detected for MMLE for τ_2 and τ_3. The ability distribution for these data was N(-1,1), indicating a non-match to the test difficulty. The results appear to suggest some lack of information may be present in a few of the data sets for estimating these category parameters. Parameters for the PCM and RSM appear to have been recovered well in every condition (see Tables 3-2 and 3-3). Recovery was nearly the same for both MMLE and MCMC for the slope and location parameters for these two models. For the GRM (see Table 3-4), the recovery of all the item parameters appear good for MCMC algorithms. Recovery for the MMLE, however, was not as good for the N(-1,1) ability distribution.

Insert Table 4-1, 4-2, 4-3, and 4-4 About Here

Model Selection Indices. In Tables 4-1 to 4-4, we present the average values of DIC, CVLL, and G^2 over the 10 replications in each condition. As described above, LR, AIC, and BIC indices are obtained based on G^2. The three values are labeled DIC-GR, CVLL-GR, G^2-GR to indicate they were estimated by calibrating the data using the GRM. Likewise, DIC-GP, CVLL-GP, G^2-GP are used to indicate the indices were obtained for the GPCM; DIC-P, CVLL-P, G^2-P indicate calibration was done using the PCM; and DIC-R, CVLL-R, G^2-R indicate calibration was done using the RSM.

In Table 4-1, the average DIC-GP values were smallest for the GPCM as the generating model is GPCM for each condition. For this same generating model, the average CVLL-GP values were larger than for other average CVLL values for each condition. Since the average G^2-GP values were smaller than for the other models for most conditions, it was reasonable to expect that the other model selection indices would also

suggest the GPCM. Similar patterns are evident in Tables 4-2 and 4-3: The average DIC and G^2 for true model was smaller than other DIC and G^2 averages, respectively, and the average CVLLs for the true model was larger than other average CVLLs for each condition. As shown in Table 4-4 when the true model is GRM, the GPCM was often selected as the better model by the DIC and PsBF indices. It appears that the power of the model selection methods may be lower when the true model was GRM.

The information in Table 5 presents the frequencies of model selection for each of the five indices for each of the conditions of the study. When data were generated by a RSM, the expectation was that a RSM would be selected as the best model.

Insert Table 5 About Here

For example, for the 10 replications in the 20-item test, $N = 1000$, and $\theta \sim N(0, 1)$ condition for data generated with RSM (see the very bottom row of Table 5), the DIC index selected GRM 0 times, GPCM 0 times, PCM 0 times and RSM 10 times as the best model. In this condition, all the five indices selected the true model. The last four columns of Table 5 present the model selection performance for the LR. Here, only the three hierarchically related models (GPCM, PCM, and RSM) were considered. LR worked very well, demonstrating about 98% ($= 1 - 9/480$) accuracy in finding the true models. AIC and BIC also demonstrated good accuracy when the generating ability distribution was N(0,1). These two indices had some difficulty in identifying the true GRM when $\theta \sim N(-1, 1)$. This appeared to be related to the poor estimation by Parscale for those conditions. This result is apparent in Figure 4 (described below). DIC functioned well for 20 items and 1,000 examinees conditions, but appeared to be less accurate in recognizing the true GRM in the other conditions. For example, in the conditions of 10 items and 500 examinees, DIC worked with only 40% ($= 1 - 16/40$)

accuracy in finding the true GRM. PsBF also performed well for selection of GPCM, PCM, and RSM in most conditions, but had some problems in finding the true GRM in more than half of the conditions.

Summaries of the performance of three indices (DIC, PsBF, and BIC) are plotted in Figures 1 to 4. LR is not shown since it was not appropriate for comparisons involving the GRM and the other models considered in this study. Further, since the performance of AIC was similar to BIC, AIC was left off the figures to enhance readability. Finally, only the GRM and GPCM were considered in these figures since all four model selection indices performed well for the PCM and RSM. The plots in Figures 1 to 4 show the proportions of model selections for the different three indices for each of the simulation conditions.

Insert Figure 1 About Here

In Figure 1, the model selection proportions are plotted between GRM and GPCM for different test length ($n = 10$, and $n = 20$). When the true model was GPCM, the three indices performed well in selection of the correct model. When the true model was the GRM, however, the DIC showed poor performance when $n = 10$. All the three indices showed moderate accuracy when $n = 20$.

Insert Figure 2 About Here

In Figure 2, the model selection proportions are plotted between GRM and GPCM 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 wrong model (i.e., the GPCM) as the better model more than half

the time. When $N = 1,000$, DIC performed better, accurately selecting the correct model about 85% of the cases. In Figure 3, the performance of the PsBF suggest it was sensitive to the number of categories when the true model was the GRM. For a test with five-category items, PsBF selected the true GRM with 90% accuracy.

Insert Figure 3 About Here

Figure 4 present model selection proportions between the GRM and GPCM for the two different ability distribution, ($N(-1, 1)$ and $N(0, 1)$). When the true model was the GRM, DIC and PsBF both demonstrated moderate accuracy in selecting the correct model. The performance BIC, however, differed depending on the ability distribution. When $\theta \sim N(0, 1)$, BIC worked perfectly.

Insert Figure 4 About Here

Discussion & Conclusions

When model selection is inaccurate, then the benefits of the model do not attach to the resulting parameter estimates. The LR test has been used often for model selection. As expected, LR showed good performance to compare hierarchically nested models in this study. When models are not nested, however, the LR test is not appropriate and other methods such as one of the other four methods discussed in this study will need to be considered.

Differences in model selection were detected among the five indices examined in this study. AIC and BIC appeared to be capable of accurately selecting the correct model for either nested or non-nested models except when the ability distribution and the test

difficulty were not matched. DIC and PsBF, which are model selection methods based on Bayesian parameter estimation, appeared to be useful for model selection when the numbers of items and examinees were large. The peformance of DIC and PsBF were inconsistent, however, for fewer items and examinees. In general, it appears that for comparisons between the GRM and GPCM, the five indices were useful, when the true model was GPCM. When the true model was the GRM, model selection accuracy for DIC and PsBF differed according to the given conditions.

Since the MCMC algorithm as implemented in Winbugs provided good recovery in item parameter estimation (as shown in Table 3), reasonable confidence was possible in the item parameter estimates used for estimating the DIC and PsBF in this study. It is likewise evident that additional conditions need to be studied to better understand how these two indices perform. They do appear to be useful, however, model selection for nested or non-nested model under the conditions simulated in this study. The MMLE algorithm as implemented in Parscale provided poor to adequate recovery for the GRM when the ability distribution did not match the difficulty of the test. The reason for the poor performances of AIC and BIC for these conditions, therefore, was not clear. Further study is suggested for the GRM calibration by Parscale. Reise and Yu (1990), in a recovery study for the GRM MULTILOG (Thissen, 1991) concluded at least 500 examinees were needed for adequate calibration of 25 items with 5 categories. They did not consider the case where ability and test difficulty were not matched.

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. Deciding which of these conditions holds for a particular test or model or set of data, however, is difficult at best, since the true model is not known for real data. Consequently, one needs to look

to the results of studies such as this one to help inform a decision as to which of the indices provides the most consistently accurate results.

In addition, it is important to consider other, non-statistical issues in a model selection process. Some models may be more appropriate for one type of psychological process than another or for one test purposes than another. This article was intended to have some contribution for the situation in which when one needs to compare various models statistically. Even though GPCM appeared to be more accurately selected by the five indices considered in this study, it may be that other reasons exist, such as the benefits of a Rasch model, and should be considered for selection of PCM or RSM.

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. (1992). *Item Response Theory.* 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., 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.

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.

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.

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 and 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.

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

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 algo-

rithm. *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.

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.

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

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.

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

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.]

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.

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

	Model Selection Methods					
	DIC	PsBF	LR test		AIC	BIC
Model		(CVLLs)	G^2	LR		
RSM	26005.10	-11950	26692.25		26704.25	26740.29
PCM	23375.70	-10625	24237.29	2545.96*	24257.29	24317.36
GPCM	22954.30	-10393	24121.30	115.99#	24151.30	24241.40
GRM	22769.70	-10292	24121.31		24151.31	24241.41

* p $(\chi^2_{df=4} > 9.49) < 0.05$
# p $(\chi^2_{df=5} > 11.07) < 0.05$

Table 2-1. Generating Item Parameters When the # of Categories is 5

	GRM					GPCM				
Item	a	b1	b2	b3	b4	a	b	τ_1	τ_2	τ_3
1	1.1886	-1.5932	-0.8280	1.2501	2.2811	1.1586	-0.4198	2.5619	-0.0422	-1.6654
2	0.9645	-2.3526	-0.2945	0.6007	1.8355	0.5127	-0.2436	0.8767	0.4519	-1.6650
3	1.5168	-0.6712	-0.0575	1.2846	2.3863	1.4293	0.6138	3.0479	-0.1036	-0.9495
4	2.4825	-1.1982	-0.0425	1.2228	2.4250	2.2519	-0.3700	-0.4075	1.8816	-0.0003
5	0.5849	-1.8364	-1.1341	-0.1715	0.6206	0.7076	0.1630	2.3457	0.1124	-0.6714
6	1.1326	-3.6806	-2.2300	-0.3021	1.4804	1.5357	0.5965	1.4453	0.0784	-0.2638
7	1.6336	-0.5844	1.0582	1.8108	2.6156	1.8720	0.1134	1.2717	-0.2400	0.4988
8	0.8229	-3.8306	-0.9803	0.4869	1.1248	0.4507	-0.4001	1.8994	-0.6004	-0.2805
9	1.9720	-3.5078	-1.2555	0.1301	0.7902	0.4865	-0.3764	3.1719	-0.0391	-2.0811
10	1.2124	-2.5060	-1.6501	0.7196	1.6200	1.3305	0.1504	1.5872	-0.1491	-0.3405
11	1.0975	-2.1523	-1.3996	0.5944	1.4757	0.8188	-0.1870	2.2012	-0.3783	-1.1993
12	0.7977	0.2119	1.1438	2.0418	2.8111	1.4120	-0.0299	0.7333	0.6034	-0.7403
13	2.0213	-3.0703	-1.1250	0.3347	1.5155	1.5035	0.3642	1.2341	1.1193	0.3773
14	1.8480	-0.6373	0.2158	1.0017	1.8338	1.4275	0.3456	0.0307	1.0191	0.2828
15	1.4760	-1.9734	-0.0262	0.9636	2.4115	1.9062	-0.2920	0.4943	1.5608	-1.3580
16	1.4036	-2.6372	-1.3044	-0.3281	0.6323	1.3970	-0.3448	1.6764	0.2668	-0.0173
17	2.4665	-2.0896	-0.9352	1.4185	2.3993	1.8138	0.1603	1.1576	0.4195	-1.2433
18	0.9345	-1.9131	-0.7949	0.4391	1.2628	0.5481	-0.2463	2.1405	-0.1835	-1.4376
19	1.2435	-1.6127	-0.6638	1.6621	2.8455	0.9901	0.2077	1.6048	-0.8582	0.4070
20	1.6544	-2.0474	-0.1620	0.6668	1.9576	0.9246	0.1946	1.6204	0.9231	-0.1622
Mean	1.4227	-1.9841	-0.6233	0.7913	1.8162	1.2239	0.0000	1.5347	0.2921	-0.6255
SD	0.5324	1.0741	0.8554	0.6807	0.7012	0.5286	0.3347	0.9224	0.7114	0.7872

Table 2-2. Generating Item Parameters When the # of Categories is 3

	GRM			GPCM		
Item	a	b1	b2	a	b	τ_1
1	1.1886	-1.2106	1.7656	1.1586	-0.4198	1.2599
2	0.9645	-1.3236	1.2181	0.5127	-0.2436	0.6643
3	1.5168	-0.3644	1.8354	1.4293	0.6138	1.4721
4	2.4825	-0.6204	1.8239	2.2519	-0.3700	0.7371
5	0.5849	-1.4853	0.2246	0.7076	0.1630	1.2291
6	1.1326	-2.9553	0.5891	1.5357	0.5965	0.7619
7	1.6336	0.2369	2.2132	1.8720	0.1134	0.5159
8	0.8229	-2.4055	0.8059	0.4507	-0.4001	0.6495
9	1.9720	-2.3817	0.4602	0.4865	-0.3764	1.5664
10	1.2124	-2.0780	1.1698	1.3305	0.1504	0.7191
11	1.0975	-1.7760	1.0351	0.8188	-0.1870	0.9115
12	0.7977	0.6778	2.4265	1.4120	-0.0299	0.6684
13	2.0213	-2.0976	0.9251	1.5035	0.3642	1.1767
14	1.8480	-0.2108	1.4178	1.4275	0.3456	0.5249
15	1.4760	-0.9998	1.6876	1.9062	-0.2920	1.0276
16	1.4036	-1.9708	0.1521	1.3970	-0.3448	0.9716
17	2.4665	-1.5124	1.9089	1.8138	0.1603	0.7886
18	0.9345	-1.3540	0.8510	0.5481	-0.2463	0.9785
19	1.2435	-1.1383	2.2538	0.9901	0.2077	0.3733
20	1.6544	-1.1047	1.3122	0.9246	0.1946	1.2718
Mean	1.4227	-1.3038	1.3038	1.2239	0.0000	0.9134
SD	0.5324	0.9203	0.6763	0.8554	0.3347	0.3306

Table 3-1. GPCM Recovery Statistics: RMSE(σ)

				MMLE				
test length	sample size	# of categ.	ability distr.	a	b	τ_1	τ_2	τ_3
n=10	500	NC=3	N(-1,1)	0.15(0.97)	0.13(0.95)	0.13(0.94)	( )	( )
			N(0,1)	0.14(0.97)	0.09(0.98)	0.18(0.90)	( )	( )
		NC=5	N(-1,1)	0.15(0.97)	0.11(0.96)	0.21(0.98)	0.20(0.96)	0.31(0.93)
			N(0,1)	0.12(0.98)	0.07(0.99)	0.24(0.97)	0.22(0.94)	0.22(0.97)
	1000	NC=3	N(-1,1)	0.12(0.98)	0.09(0.97)	0.13(0.94)	( )	( )
			N(0,1)	0.10(0.98)	0.07(0.98)	0.12(0.96)	( )	( )
		NC=5	N(-1,1)	0.22(0.93)	0.12(0.96)	0.23(0.98)	1.96(0.26)	2.19(0.46)
			N(0,1)	0.08(0.99)	0.05(0.99)	0.17(0.99)	0.16(0.97)	0.17(0.98)
n=20	500	NC=3	N(-1,1)	0.14(0.97)	0.11(0.95)	0.14(0.93)	( )	( )
			N(0,1)	0.12(0.97)	0.09(0.97)	0.12(0.94)	( )	( )
		NC=5	N(-1,1)	0.10(0.98)	0.10(0.96)	0.18(0.98)	0.19(0.96)	0.25(0.95)
			N(0,1)	0.12(0.98)	0.07(0.98)	0.23(0.97)	0.18(0.97)	0.19(0.97)
	1000	NC=3	N(-1,1)	0.10(0.98)	0.08(0.97)	0.09(0.96)	( )	( )
			N(0,1)	0.09(0.99)	0.06(0.99)	0.10(0.96)	( )	( )
		NC=5	N(-1,1)	0.09(0.99)	0.06(0.98)	0.15(0.99)	0.13(0.98)	0.19(0.97)
			N(0,1)	0.07(0.99)	0.04(0.99)	0.15(0.99)	0.13(0.98)	0.14(0.98)
				MCMC				
test length	sample size	# of categ.	ability distr.	a	b	τ_1	τ_2	τ_3
n=10	500	NC=3	N(-1,1)	0.15(0.97)	0.13(0.95)	0.14(0.94)	( )	( )
			N(0,1)	0.14(0.97)	0.09(0.98)	0.19(0.89)	( )	( )
		NC=5	N(-1,1)	0.15(0.97)	0.14(0.94)	0.24(0.98)	0.22(0.95)	0.33(0.93)
			N(0,1)	0.12(0.98)	0.07(0.99)	0.24(0.98)	0.22(0.95)	0.22(0.97)
	1000	NC=3	N(-1,1)	0.12(0.98)	0.09(0.97)	0.13(0.94)	( )	( )
			N(0,1)	0.12(0.98)	0.09(0.97)	0.13(0.94)	( )	( )
		NC=5	N(-1,1)	0.10(0.99)	0.08(0.98)	0.19(0.99)	0.15(0.98)	0.19(0.97)
			N(0,1)	0.08(0.99)	0.05(0.99)	0.18(0.99)	0.16(0.97)	0.18(0.98)
n=20	500	NC=3	N(-1,1)	0.14(0.97)	0.11(0.94)	0.15(0.92)	( )	( )
			N(0,1)	0.12(0.97)	0.09(0.96)	0.12(0.93)	( )	( )
		NC=5	N(-1,1)	0.10(0.98)	0.12(0.95)	0.19(0.98)	0.21(0.96)	0.27(0.95)
			N(0,1)	0.12(0.98)	0.07(0.98)	0.22(0.97)	0.18(0.97)	0.19(0.97)
	1000	NC=3	N(-1,1)	0.10(0.98)	0.08(0.97)	0.10(0.96)	( )	( )
			N(0,1)	0.09(0.99)	0.06(0.99)	0.10(0.96)	( )	( )
		NC=5	N(-1,1)	0.09(0.99)	0.10(0.97)	0.17(0.98)	0.15(0.98)	0.21(0.97)
			N(0,1)	0.07(0.99)	0.04(0.99)	0.16(0.99)	0.13(0.98)	0.14(0.98)

Table 3-2. PCM Recovery Statistics: RMSE(σ)

				MMLE			
test length	sample size	# of categ.	ability distr.	b	τ_1	τ_2	τ_3
n=10	500	NC=3	N(-1,1)	0.09(0.97)	0.11(0.96)	( )	( )
			N(0,1)	0.07(0.98)	0.10(0.97)	( )	( )
		NC=5	N(-1,1)	0.09(0.97)	0.14(0.99)	0.15(0.98)	0.20(0.97)
			N(0,1)	0.06(0.99)	0.20(0.99)	0.17(0.98)	0.17(0.98)
	1000	NC=3	N(-1,1)	0.06(0.99)	0.09(0.97)	( )	( )
			N(0,1)	0.05(0.99)	0.07(0.98)	( )	( )
		NC=5	N(-1,1)	0.05(0.99)	0.09(1.00)	0.11(0.99)	0.16(0.98)
			N(0,1)	0.05(0.99)	0.13(0.99)	0.10(0.99)	0.12(0.99)
n=20	500	NC=3	N(-1,1)	0.09(0.97)	0.12(0.94)	( )	( )
			N(0,1)	0.07(0.98)	0.09(0.96)	( )	( )
		NC=5	N(-1,1)	0.09(0.97)	0.14(0.99)	0.16(0.98)	0.21(0.97)
			N(0,1)	0.07(0.98)	0.17(0.98)	0.15(0.98)	0.16(0.98)
	1000	NC=3	N(-1,1)	0.06(0.98)	0.07(0.98)	( )	( )
			N(0,1)	0.05(0.99)	0.07(0.98)	( )	( )
		NC=5	N(-1,1)	0.07(0.98)	0.10(0.99)	0.13(0.98)	0.16(0.98)
			N(0,1)	0.05(0.99)	0.12(0.99)	0.11(0.99)	0.10(0.99)
				MCMC			
test length	sample size	# of categ.	ability distr.	b	τ_1	τ_2	τ_3
n=10	500	NC=3	N(-1,1)	0.09(0.97)	0.11(0.96)	( )	( )
			N(0,1)	0.07(0.98)	0.10(0.97)	( )	( )
		NC=5	N(-1,1)	0.09(0.97)	0.14(0.99)	0.15(0.98)	0.20(0.97)
			N(0,1)	0.06(0.99)	0.20(0.99)	0.17(0.98)	0.17(0.98)
	1000	NC=3	N(-1,1)	0.06(0.99)	0.09(0.97)	( )	( )
			N(0,1)	0.05(0.99)	0.07(0.98)	( )	( )
		NC=5	N(-1,1)	0.05(0.99)	0.09(1.00)	0.11(0.99)	0.16(0.98)
			N(0,1)	0.04(0.99)	0.13(0.99)	0.10(0.99)	0.12(0.99)
n=20	500	NC=3	N(-1,1)	0.09(0.97)	0.12(0.94)	( )	( )
			N(0,1)	0.07(0.98)	0.09(0.96)	( )	( )
		NC=5	N(-1,1)	0.11(0.95)	0.16(0.99)	0.18(0.97)	0.22(0.96)
			N(0,1)	0.07(0.98)	0.16(0.98)	0.15(0.98)	0.16(0.98)
	1000	NC=3	N(-1,1)	0.06(0.98)	0.07(0.98)	( )	( )
			N(0,1)	0.05(0.99)	0.07(0.98)	( )	( )
		NC=5	N(-1,1)	0.07(0.98)	0.10(0.99)	0.13(0.98)	0.16(0.98)
			N(0,1)	0.05(0.99)	0.12(0.99)	0.11(0.99)	0.10(0.99)

Table 3-3. RSM Recovery Statistics: RMSE(σ)

				MMLE			
test length	sample size	# of categ.	ability distr.	b	τ_1	τ_2	τ_3
n=10	500	NC=3	N(-1,1)	0.07(0.98)	0.03()	()	()
			N(0,1)	0.07(0.98)	0.04()	()	()
		NC=5	N(-1,1)	0.06(0.99)	0.06()	0.05()	0.06()
			N(0,1)	0.05(0.99)	0.06()	0.04()	0.05()
	1000	NC=3	N(-1,1)	0.05(0.99)	0.03()	()	()
			N(0,1)	0.05(0.99)	0.03()	()	()
		NC=5	N(-1,1)	0.04(0.99)	0.04()	0.04()	0.05()
			N(0,1)	0.03(1.00)	0.04()	0.04()	0.03()
n=20	500	NC=3	N(-1,1)	0.08(0.97)	0.03()	()	()
			N(0,1)	0.07(0.98)	0.03()	()	()
		NC=5	N(-1,1)	0.07(0.98)	0.04()	0.02()	0.07()
			N(0,1)	0.05(0.99)	0.04()	0.02()	0.02()
	1000	NC=3	N(-1,1)	0.06(0.99)	0.02()	()	()
			N(0,1)	0.05(0.99)	0.02()	()	()
		NC=5	N(-1,1)	0.04(0.99)	0.03()	0.02()	0.03()
			N(0,1)	0.04(0.99)	0.04()	0.02()	0.02()
				MCMC			
test length	sample size	# of categ.	ability distr.	b	τ_1	τ_2	τ_3
n=10	500	NC=3	N(-1,1)	0.07(0.98)	0.03()	()	()
			N(0,1)	0.08(0.98)	0.04()	()	()
		NC=5	N(-1,1)	0.06(0.99)	0.06()	0.05()	0.06()
			N(0,1)	0.05(0.99)	0.06()	0.04()	0.05()
	1000	NC=3	N(-1,1)	0.05(0.99)	0.03()	()	()
			N(0,1)	0.05(0.99)	0.03()	()	()
		NC=5	N(-1,1)	0.04(0.99)	0.04()	0.04()	0.05()
			N(0,1)	0.03(1.00)	0.04()	0.04()	0.03()
n=20	500	NC=3	N(-1,1)	0.08(0.97)	0.02()	()	()
			N(0,1)	0.07(0.98)	0.03()	()	()
		NC=5	N(-1,1)	0.07(0.98)	0.04()	0.02()	0.07()
			N(0,1)	0.05(0.99)	0.04()	0.02()	0.02()
	1000	NC=3	N(-1,1)	0.06(0.99)	0.02()	()	()
			N(0,1)	0.05(0.99)	0.02()	()	()
		NC=5	N(-1,1)	0.04(0.99)	0.04()	0.02()	0.03()
			N(0,1)	0.04(0.99)	0.03()	0.02()	0.02()

Table 3-4. GRM Recovery Statistics: RMSE(σ)

				MMLE				
test length	sample size	# of categ.	ability distr.	a	b1	b2	b3	b4
n=10	500	NC=3	N(-1,1)	0.21(0.94)	0.14(0.99)	0.22(0.95)	( )	( )
			N(0,1)	0.18(0.95)	0.21(0.98)	0.15(0.98)	( )	( )
		NC=5	N(-1,1)	0.28(0.88)	0.61(0.86)	0.42(0.88)	0.45(0.76)	0.60(0.70)
			N(0,1)	0.13(0.97)	0.30(0.97)	0.12(0.99)	0.11(0.99)	0.18(0.97)
	1000	NC=3	N(-1,1)	0.35(0.83)	0.64(0.76)	0.60(0.58)	( )	( )
			N(0,1)	0.11(0.98)	0.14(0.99)	0.10(0.99)	( )	( )
		NC=5	N(-1,1)	0.33(0.85)	0.69(0.81)	0.51(0.82)	0.54(0.65)	0.67(0.56)
			N(0,1)	0.10(0.98)	0.23(0.98)	0.09(0.99)	0.08(0.99)	0.15(0.98)
n=20	500	NC=3	N(-1,1)	0.34(0.85)	0.66(0.77)	0.68(0.62)	( )	( )
			N(0,1)	0.16(0.96)	0.19(0.98)	0.17(0.97)	( )	( )
		NC=5	N(-1,1)	0.21(0.93)	0.62(0.84)	0.39(0.90)	0.30(0.91)	0.41(0.86)
			N(0,1)	0.11(0.98)	0.22(0.98)	0.13(0.99)	0.13(0.98)	0.17(0.97)
	1000	NC=3	N(-1,1)	0.32(0.85)	0.58(0.77)	0.59(0.62)	( )	( )
			N(0,1)	0.11(0.98)	0.12(0.99)	0.12(0.99)	( )	( )
		NC=5	N(-1,1)	0.28(0.88)	0.62(0.82)	0.43(0.86)	0.44(0.78)	0.56(0.72)
			N(0,1)	0.08(0.99)	0.19(0.99)	0.09(0.99)	0.09(0.99)	0.12(0.98)
				MCMC				
test length	sample size	# of categ.	ability distr.	a	b1	b2	b3	b4
n=10	500	NC=3	N(-1,1)	0.21(0.94)	0.16(0.99)	0.24(0.94)	( )	( )
			N(0,1)	0.18(0.95)	0.24(0.97)	0.16(0.97)	( )	( )
		NC=5	N(-1,1)	0.13(0.97)	0.21(0.98)	0.13(0.99)	0.17(0.97)	0.40(0.89)
			N(0,1)	0.13(0.97)	0.41(0.95)	0.15(0.99)	0.12(0.99)	0.20(0.96)
	1000	NC=3	N(-1,1)	0.12(0.98)	0.11(0.99)	0.18(0.96)	( )	( )
			N(0,1)	0.11(0.98)	0.15(0.99)	0.11(0.99)	( )	( )
		NC=5	N(-1,1)	0.09(0.99)	0.14(0.99)	0.09(1.00)	0.14(0.98)	0.37(0.88)
			N(0,1)	0.10(0.98)	0.24(0.98)	0.09(0.99)	0.08(0.99)	0.15(0.98)
n=20	500	NC=3	N(-1,1)	0.18(0.95)	0.16(0.99)	0.26(0.94)	( )	( )
			N(0,1)	0.15(0.96)	0.21(0.98)	0.18(0.97)	( )	( )
		NC=5	N(-1,1)	0.13(0.97)	0.20(0.99)	0.11(0.99)	0.20(0.96)	0.38(0.90)
			N(0,1)	0.12(0.98)	0.24(0.98)	0.13(0.99)	0.13(0.98)	0.19(0.97)
	1000	NC=3	N(-1,1)	0.10(0.98)	0.09(1.00)	0.17(0.97)	( )	( )
			N(0,1)	0.11(0.98)	0.12(0.99)	0.12(0.99)	( )	( )
		NC=5	N(-1,1)	0.09(0.99)	0.10(1.00)	0.08(1.00)	0.13(0.98)	0.22(0.96)
			N(0,1)	0.08(0.99)	0.20(0.98)	0.10(0.99)	0.09(0.99)	0.13(0.98)

Table 4-1. Averages of Model Selection Indices When Data Generated With GPCM

test length	sample size	# of categ.	ability distr.	Average indices of model-selection methods (SD) DIC-GR	DIC-GP	DIC-P	DIC-R	CVLL-GR	CVLL-GP
n=10	500	NC=3	N(-1,1)	8227.83	8210.47	8456.61	8559.12	-3878.33	-3871.92
				(47.77)	(47.83)	(63.66)	(69.21)	(7.94)	(6.96)
			N(0,1)	8958.86	8945.76	9240.11	9391.02	-4251.13	-4248.10
				(71.10)	(70.16)	(59.99)	(47.37)	(10.03)	(10.16)
		NC=5	N(-1,1)	10868.06	10759.84	11123.36	12041.41	-5280.38	-5238.14
				(155.21)	(153.79)	(146.91)	(111.67)	(19.45)	(16.13)
			N(0,1)	12105.96	12000.60	12418.10	13168.42	-5771.53	-5740.22
				(58.15)	(57.34)	(81.46)	(81.37)	(9.70)	(7.77)
	1000	NC=3	N(-1,1)	16200.15	16179.43	16717.81	16943.11	-7782.64	-7744.80
				(151.44)	(151.75)	(137.79)	(149.37)	(53.61)	(15.27)
			N(0,1)	17718.57	17700.16	18291.38	18591.77	-8454.62	-8459.27
				(106.87)	(107.23)	(88.57)	(88.07)	(12.41)	(16.62)
		NC=5	N(-1,1)	21603.17	21419.66	22144.68	24039.32	-10463.84	-10397.03
				(135.74)	(134.13)	(126.71)	(156.67)	(10.78)	(12.94)
			N(0,1)	23992.20	23834.21	24659.04	26189.41	-11534.71	-11467.75
				(220.90)	(212.52)	(167.39)	(162.43)	(17.00)	(14.56)
n=20	500	NC=3	N(-1,1)	15578.10	15529.12	15944.91	16111.66	-7556.74	-7541.39
				(235.23)	(236.35)	(237.13)	(251.60)	(14.42)	(12.10)
			N(0,1)	17331.42	17296.55	17747.14	18007.60	-8399.48	-8385.12
				(180.53)	(182.50)	(139.15)	(144.29)	(12.12)	(13.38)
		NC=5	N(-1,1)	21066.99	20838.91	21452.04	23200.29	-10244.96	-10178.94
				(283.85)	(271.21)	(253.77)	(291.01)	(8.62)	(7.26)
			N(0,1)	23243.91	23026.14	23701.74	25539.18	-11212.89	-11146.44
				(190.90)	(181.48)	(194.26)	(197.52)	(19.61)	(17.51)
	1000	NC=3	N(-1,1)	31273.87	31217.71	32103.78	32467.20	-15151.43	-15133.44
				(287.62)	(284.84)	(236.48)	(231.09)	(10.40)	(8.24)
			N(0,1)	34470.82	34420.76	35343.29	35922.71	-16850.19	-16834.62
				(163.73)	(169.46)	(168.18)	(155.17)	(14.51)	(13.42)
		NC=5	N(-1,1)	41743.76	41417.15	42649.15	46309.79	-20315.48	-20227.88
				(375.87)	(392.83)	(399.35)	(371.08)	(15.66)	(16.15)
			N(0,1)	46525.63	46205.39	47552.28	51319.52	-22709.24	-22592.41
				(318.29)	(318.98)	(267.80)	(269.75)	(13.58)	(14.82)

test length	sample size	# of categ.	ability distr.	CVLL-P	CVLL-R	G^2-GR	G^2-GP	G^2-P	G^2-R
n=10	500	NC=3	N(-1,1)	-3986.99	-4041.36	8559.73	8551.62	8725.87	8844.01
				(2.52)	(23.80)	(50.05)	(48.86)	(57.93)	(64.36)
			N(0,1)	-4408.33	-4479.00	9343.30	9337.74	9552.69	9719.28
				(6.57)	(4.77)	(58.83)	(58.54)	(56.13)	(43.01)
		NC=5	N(-1,1)	-5427.02	-5825.37	11338.67	11292.35	11780.06	13225.02
				(10.16)	(3.71)	(138.27)	(133.73)	(699.36)	(1231.92)
			N(0,1)	-5952.46	-6353.54	12657.38	12610.65	12920.28	13716.64
				(6.94)	(4.96)	(48.17)	(47.58)	(61.40)	(79.49)
	1000	NC=3	N(-1,1)	-8021.11	-8127.87	16960.35	16949.44	17326.99	17566.87
				(6.32)	(3.44)	(127.83)	(126.49)	(120.74)	(131.82)
			N(0,1)	-8759.74	-8907.06	18592.93	18582.64	18993.68	19307.02
				(5.37)	(3.60)	(69.30)	(70.40)	(71.89)	(70.88)
		NC=5	N(-1,1)	-10808.24	-11703.25	22656.65	22680.76	22949.53	26190.14
				(6.59)	(4.50)	(121.77)	(403.86)	(221.58)	(2330.62)
			N(0,1)	-11901.23	-12654.90	25250.93	25166.59	25780.16	27395.79
				(11.13)	(5.13)	(167.60)	(160.02)	(145.84)	(125.79)
n=20	500	NC=3	N(-1,1)	-7704.92	-7807.92	16106.83	16079.72	16443.74	16801.69
				(10.46)	(7.07)	(236.72)	(239.19)	(251.56)	(455.97)
			N(0,1)	-8643.95	-8770.32	17935.00	17919.44	18323.33	18618.07
				(6.72)	(6.54)	(145.82)	(146.87)	(111.21)	(116.47)
		NC=5	N(-1,1)	-10507.95	-11412.93	21714.83	21585.50	21836.88	26708.87
				(9.54)	(2.89)	(269.22)	(258.44)	(260.15)	(4070.54)
			N(0,1)	-11482.17	-12485.09	24007.09	23884.09	24478.90	26593.08
				(13.33)	(4.53)	(163.30)	(156.67)	(177.14)	(523.60)
	1000	NC=3	N(-1,1)	-15561.67	-15752.03	32456.22	32421.83	33175.14	33572.52
				(5.60)	(5.50)	(270.66)	(269.55)	(227.39)	(223.00)
			N(0,1)	-17281.16	-17560.05	35831.56	35800.55	36586.61	37195.95
				(10.21)	(5.32)	(144.27)	(147.99)	(148.07)	(132.07)
		NC=5	N(-1,1)	-20839.26	-22599.45	43344.61	43123.94	44708.88	49083.12
				(9.49)	(4.38)	(366.09)	(381.75)	(3922.05)	(3232.45)
			N(0,1)	-23320.61	-25150.01	48329.60	48105.44	49179.56	53147.16
				(10.93)	(5.82)	(268.37)	(274.87)	(276.72)	(218.78)

Table 4-2. Averages of Model Selection Indices When Data Generated With PCM

test length	sample size	# of categ.	ability distr.	Average indices of model-selection methods (SD)					
				DIC-GR	DIC-GP	DIC-P	DIC-R	CVLL-GR	CVLL-GP
n=10	500	NC=3	N(-1,1)	8459.81	8440.44	8434	8537.05	-4113.17	-4100.35
				(124.83)	(125.09)	(121.51)	(111.45)	(11.54)	(10.77)
			N(0,1)	9138.36	9123	9115.32	9255.64	-4308.87	-4302.59
				(89.59)	(90.51)	(87.97)	(80.94)	(16.42)	(15.51)
		NC=5	N(-1,1)	11066.35	10968.59	10955.41	11726.14	-5332.65	-5302.91
				(186.23)	(184.86)	(181.87)	(191.89)	(16.47)	(8.27)
			N(0,1)	12262.52	12173.91	12167.34	12861.2	-5924	-5899.1
				(132.5)	(124.78)	(119.07)	(104.87)	(11.04)	(10.86)
	1000	NC=3	N(-1,1)	16925.48	16898.9	16895.86	17119.74	-7980.45	-7972.27
				(172.32)	(168.22)	(163.36)	(156.82)	(12.16)	(13.18)
			N(0,1)	18272.63	18258.08	18250.99	18552.38	-8689.78	-8673.29
				(131.72)	(133.56)	(129.03)	(114.39)	(14.88)	(12.99)
		NC=5	N(-1,1)	22059.61	21893.8	21884.15	23477.55	-10619.21	-10556.65
				(111.42)	(110.96)	(110.36)	(106.33)	(8.71)	(9.85)
			N(0,1)	24544.49	24417.25	24401.11	25765.91	-11708.74	-11661.05
				(144.5)	(151.76)	(147.41)	(91.1)	(7.47)	(11.59)
n=20	500	NC=3	N(-1,1)	16681.73	16637.14	16616.14	16791.55	-8020.71	-8011.02
				(161.03)	(167.24)	(164.04)	(148.37)	(10.94)	(11.44)
			N(0,1)	18220.74	18184.72	18159.1	18368.24	-8670.21	-8663.99
				(115.67)	(115.02)	(108.19)	(115.09)	(7.6)	(8.41)
		NC=5	N(-1,1)	22041.61	21843.23	21820.66	23263.55	-10783.27	-10756.36
				(130.99)	(128.04)	(128.85)	(152.12)	(19.04)	(19.6)
			N(0,1)	24564.5	24389.87	24368.98	25767.26	-12173.95	-12129.79
				(153.75)	(160.36)	(154.25)	(139.87)	(13.1)	(15.59)
	1000	NC=3	N(-1,1)	33069.23	33006.38	32992.26	33346.34	-16141.22	-16121.02
				(302.22)	(303.14)	(297.26)	(310.11)	(14.76)	(14.61)
			N(0,1)	36124.18	36067.9	36047.7	36499.86	-17414.76	-17383.76
				(178.18)	(187.03)	(187.57)	(201.68)	(22.98)	(11.72)
		NC=5	N(-1,1)	44016.36	43716.58	43691.77	46650.29	-21644.81	-21552.78
				(459.96)	(457.51)	(455.74)	(455.37)	(13.15)	(15.65)
			N(0,1)	48837.64	48557.13	48532.01	51446.77	-23924.22	-23805.12
				(259.85)	(254.46)	(251.32)	(233.93)	(22.2)	(22.08)

test length	sample size	# of categ.	ability distr.	CVLL-P	CVLL-R	G^2-GR	G^2-GP	G^2-P	G^2-R
n=10	500	NC=3	N(-1,1)	-4089.16	-4154.71	8726.07	8716.79	8726.53	8845.49
				(6.8)	(2.93)	(114.93)	(114.88)	(113.85)	(105.32)
			N(0,1)	-4294.73	-4377.04	9446.25	9439.07	9448.22	9604
				(11.44)	(9.25)	(66.04)	(66.86)	(68.48)	(63.78)
		NC=5	N(-1,1)	-5293.54	-5667.57	11475.18	11421.78	11430.29	12238.25
				(6.55)	(3.59)	(165.89)	(166.66)	(168.42)	(180.35)
			N(0,1)	-5890.14	-6214.61	12761.86	12717.41	12730.84	13466.79
				(8.89)	(5.09)	(109.19)	(102.75)	(101.56)	(85.49)
	1000	NC=3	N(-1,1)	-7967.14	-8123.81	17538.92	17523.4	17534.45	17772.42
				(5.88)	(5.13)	(155.38)	(152.24)	(152.18)	(142.51)
			N(0,1)	-8664.51	-8803.31	18953.95	18946.03	18956.25	19270.85
				(5.43)	(5.47)	(106.22)	(108.98)	(110.08)	(90.41)
		NC=5	N(-1,1)	-10554.52	-11350.73	23016.98	22899.36	22908.13	24523.65
				(6.1)	(5.04)	(102.18)	(99.64)	(101.85)	(98.38)
			N(0,1)	-11650.61	-12330.55	25651.33	25569.09	25578.42	26975.77
				(7.07)	(6.53)	(135.07)	(138.2)	(139.06)	(87.31)
n=20	500	NC=3	N(-1,1)	-7995.65	-8083.48	17108.86	17083.9	17102.56	17313.44
				(6.59)	(4.38)	(150.00)	(154.38)	(152.92)	(137.25)
			N(0,1)	-8649.3	-8747.97	18699.74	18680.79	18695.82	18940.69
				(7.83)	(4.23)	(100.44)	(100.05)	(96.07)	(104.89)
		NC=5	N(-1,1)	-10737.11	-11374.96	22611.4	22493.27	22513.97	24052.77
				(14.09)	(3.29)	(117.89)	(113.49)	(115.79)	(141.95)
			N(0,1)	-12113.26	-12808.89	25215.88	25117.54	25141.54	26642.51
				(15.67)	(6.58)	(133.88)	(142.88)	(139.59)	(125.54)
	1000	NC=3	N(-1,1)	-16110.53	-16276.95	34085.35	34041.83	34064.54	34452.27
				(6.35)	(5.32)	(261.76)	(263.08)	(260.82)	(275.00)
			N(0,1)	-17373.16	-17606.67	37244.14	37210.02	37227.27	37712.15
				(9.31)	(9.36)	(157.47)	(164.04)	(165.9)	(181.03)
		NC=5	N(-1,1)	-21538.92	-22973.43	45409.01	45196.04	45215.7	48257.15
				(10.85)	(4.62)	(443.85)	(442.19)	(442.75)	(446.94)
			N(0,1)	-23788.98	-25220.09	50420.79	50218.58	50238.88	53244.24
				(19.54)	(7.7)	(236.13)	(228.28)	(228.97)	(211.71)

Table 4-3. Averages of Model Selection Indices When Data Generated With RSM

test length	sample size	# of categ.	ability distr.	Average indices of model-selection methods (SD) DIC-GR	DIC-GP	DIC-P	DIC-R	CVLL-GR	CVLL-GP
n=10	500	NC=3	N(-1,1)	8434.44	8419.28	8413.89	8404.01	-3990.68	-3982.19
				(115.24)	(117.71)	(116.3)	(115.67)	(16.3)	(14.82)
			N(0,1)	8915.39	8905.25	8905.09	8895.96	-4217.29	-4210.13
				(97.11)	(96.04)	(86.56)	(87.63)	(11.47)	(11.88)
		NC=5	N(-1,1)	10528.71	10446.74	10438.67	10413.58	-4981.55	-4961.71
				(80.47)	(78.52)	(75.99)	(69.9)	(10.49)	(11.46)
			N(0,1)	11741.14	11652.05	11640.9	11611.17	-5607.94	-5584.7
				(85.26)	(89.29)	(86.01)	(85.55)	(10.11)	(8.65)
	1000	NC=3	N(-1,1)	16800.37	16777.92	16768.85	16758.59	-8057.44	-8031.59
				(161.07)	(161.14)	(156.16)	(155.85)	(30.16)	(11.43)
			N(0,1)	17745.77	17728.33	17723.65	17715.59	-8541.04	-8518.61
				(176.82)	(175.63)	(168.19)	(167.11)	(29.32)	(11.61)
		NC=5	N(-1,1)	21107.86	20978.89	20967	20944.77	-10155.45	-10117.4
				(161.11)	(167.62)	(164.47)	(166.72)	(13.56)	(11.99)
			N(0,1)	23427.45	23303.39	23294.36	23269.42	-11254.11	-11212.23
				(162.54)	(166.33)	(161.25)	(163.55)	(10.00)	(9.18)
n=20	500	NC=3	N(-1,1)	16542.21	16511.37	16496.67	16478.01	-7951.24	-7946.25
				(143.25)	(145.67)	(141.90)	(143.28)	(14.88)	(11.69)
			N(0,1)	17605.12	17577.02	17555.2	17536.53	-8470.58	-8464.08
				(143.12)	(142.23)	(137.77)	(141.55)	(8.25)	(8.49)
		NC=5	N(-1,1)	20849.53	20673.24	20658.04	20604.4	-10123.09	-10060.06
				(181.57)	(183.25)	(174.74)	(177.74)	(14.06)	(13.05)
			N(0,1)	23182.65	23012.52	22989.44	22932.16	-11356.96	-11318.59
				(163.96)	(158.43)	(154.32)	(158.3)	(14.33)	(14.21)
	1000	NC=3	N(-1,1)	32905.13	32856.22	32837.89	32821.44	-15997.95	-15917.28
				(209.15)	(218.23)	(215.38)	(217.55)	(58.78)	(14.30)
			N(0,1)	34960.39	34920.87	34905.22	34884.97	-17113.14	-17079.3
				(193.14)	(199.67)	(195.71)	(194.83)	(16.75)	(4.92)
		NC=5	N(-1,1)	41309.91	41060.98	41044.95	40996.11	-20101.45	-20017.91
				(265.95)	(273.12)	(273.64)	(269.51)	(14.83)	(14.73)
			N(0,1)	46184.22	45923.3	45901.16	45846.55	-22866.9	-22772.79
				(277.60)	(281.59)	(275.71)	(278.48)	(8.76)	(8.99)

test length	sample size	# of categ.	ability distr.	CVLL-P	CVLL-R	G^2-GR	G^2-GP	G^2-P	G^2-R
n=10	500	NC=3	N(-1,1)	-3974.36	-3970.17	8687.45	8681.14	8690.91	8698.65
				(2.58)	(2.28)	(101.74)	(103.63)	(104.7)	(104.22)
			N(0,1)	-4202.34	-4197.57	9213.84	9209.38	9221.51	9230.27
				(6.9)	(4.31)	(86.57)	(86.18)	(81.49)	(82.69)
		NC=5	N(-1,1)	-4954.07	-4939.76	10859.65	10816.32	10826.01	10853.42
				(11.16)	(5.01)	(84.51)	(83.34)	(81.46)	(75.96)
			N(0,1)	-5576.75	-5565.05	12179.62	12130.57	12138.99	12162.9
				(7.35)	(3.18)	(85.38)	(90.19)	(88.68)	(87.69)
	1000	NC=3	N(-1,1)	-8020.5	-8016.63	17357.94	17345.5	17354.06	17361.64
				(4.32)	(4.01)	(144.8)	(144.23)	(143.98)	(143.74)
			N(0,1)	-8516.25	-8510.96	18382.89	18373.95	18383.48	18392.89
				(6.03)	(5.11)	(141.67)	(140.46)	(139.99)	(139.38)
		NC=5	N(-1,1)	-10108.62	-10092.85	21887.19	21800.78	21810.63	21840.6
				(11.32)	(5.35)	(148.51)	(159.29)	(159.36)	(161.89)
			N(0,1)	-11209.33	-11196.48	24450.44	24364.8	24374.19	24403.1
				(5.1)	(6.36)	(152.9)	(159.24)	(158.07)	(160.55)
n=20	500	NC=3	N(-1,1)	-7926.92	-7918.63	16949.13	16935.47	16957.45	16976.18
				(8.4)	(6.38)	(126.91)	(128.13)	(126.51)	(127.77)
			N(0,1)	-8451.41	-8442.09	18053.53	18040.18	18056.28	18075.76
				(9.62)	(8.08)	(128.97)	(128.36)	(126.34)	(129.80)
		NC=5	N(-1,1)	-10040.37	-10011.27	21354.79	21280.89	21227.66	21286.14
				(9.42)	(4.26)	(217.93)	(234.8)	(172.82)	(176.21)
			N(0,1)	-11299.62	-11273.25	23748.16	23654.78	23674.9	23730.7
				(9.72)	(4.23)	(142.55)	(138.23)	(136.73)	(141.26)
	1000	NC=3	N(-1,1)	-15904.14	-15893.87	33865.74	33837.44	33856.05	33877.33
				(9.65)	(6.02)	(171.06)	(177.1)	(179.00)	(181.36)
			N(0,1)	-17067.27	-17057.84	35995.08	35972.67	35992.31	36009.82
				(2.33)	(4.55)	(169.34)	(174.94)	(173.70)	(172.79)
		NC=5	N(-1,1)	-20004.05	-19968.96	42515.37	42341.5	42362.16	42423.9
				(9.88)	(6.2)	(261.15)	(270.85)	(271.30)	(268.51)
			N(0,1)	-22761.12	-22732.49	47602.81	47414.45	47434.62	47493.33
				(6.72)	(7.18)	(246.05)	(254.04)	(252.52)	(256.24)

Table 4-4. Averages of Model Selection Indices When Data Generated With GRM

test length	sample size	# of categ.	ability distr.	Average indices of model-selection methods (SD)					
				DIC-GR	DIC-GP	DIC-P	DIC-R	CVLL-GR	CVLL-GP
n=10	500	NC=3	N(-1,1)	7463.59	7458.77	7678.84	8038.15	-3551.21	-3544.81
				(68.25)	(68.05)	(56.08)	(76.49)	(19.58)	(12.29)
			N(0,1)	8041.64	8039.46	8270.76	8620.43	-3831.64	-3833.92
				(42.17)	(43.18)	(62.95)	(62.98)	(8.47)	(6.24)
		NC=5	N(-1,1)	11741.52	11749.29	12081.77	13000.83	-5677.59	-5698.71
				(123.43)	(125.94)	(119.85)	(107.32)	(10.44)	(11.53)
			N(0,1)	12633.88	12622.79	12993.53	13861.30	-6136.67	-6131.28
				(87.85)	(95.51)	(105.26)	(73.43)	(9.90)	(11.60)
	1000	NC=3	N(-1,1)	15052.80	15046.56	15460.75	16205.45	-7180.39	-7108.01
				(186.07)	(183.79)	(167.67)	(165.48)	(59.88)	(15.84)
			N(0,1)	16131.58	16138.54	16605.17	17311.57	-7633.38	-7628.11
				(101.85)	(100.75)	(99.87)	(97.88)	(24.63)	(12.43)
		NC=5	N(-1,1)	23170.61	23206.14	23951.37	25850.25	-11130.19	-11160.53
				(157.28)	(156.09)	(168.19)	(198.68)	(19.80)	(19.06)
			N(0,1)	25229.57	25266.00	25942.25	27726.34	-12127.45	-12183.66
				(107.33)	(104.79)	(152.34)	(161.74)	(17.51)	(21.21)
n=20	500	NC=3	N(-1,1)	14288.18	14281.78	14692.77	15418.60	-6863.96	-6866.29
				(132.24)	(138.36)	(99.68)	(102.39)	(12.77)	(8.39)
			N(0,1)	15308.97	15292.83	15730.01	16540.38	-7478.51	-7484.31
				(54.54)	(62.53)	(92.37)	(76.62)	(16.13)	(11.36)
		NC=5	N(-1,1)	22440.25	22452.92	23149.29	24961.36	-10940.21	-10991.69
				(307.10)	(317.25)	(351.17)	(350.70)	(12.65)	(13.49)
			N(0,1)	24476.55	24477.61	25185.01	27016.32	-11899.24	-11956.49
				(144.82)	(129.88)	(157.70)	(92.16)	(10.53)	(13.49)
	1000	NC=3	N(-1,1)	28548.35	28564.41	29363.68	30787.13	-13998.61	-13960.45
				(225.02)	(229.47)	(203.95)	(219.86)	(43.50)	(10.38)
			N(0,1)	30670.69	30685.22	31544.71	33246.65	-14854.73	-14849.72
				(115.02)	(113.97)	(139.15)	(161.21)	(19.22)	(10.23)
		NC=5	N(-1,1)	45134.11	45248.17	46652.08	50268.81	-22070.59	-22147.44
				(164.23)	(165.20)	(193.05)	(199.84)	(10.34)	(11.56)
			N(0,1)	48725.88	48805.10	50275.81	54025.64	-24128.60	-24210.87
				(260.55)	(274.39)	(298.85)	(235.93)	(15.35)	(15.99)

test length	sample size	# of categ.	ability distr.	CVLL-P	CVLL-R	G^2-GR	G^2-GP	G^2-P	G^2-R
n=10	500	NC=3	N(-1,1)	-3646.96	-3841.72	7776.65	7780.75	7940.33	8310.62
				(6.19)	(3.18)	(61.16)	(61.60)	(55.59)	(71.94)
			N(0,1)	-3947.88	-4112.08	8400.35	8404.52	8572.75	8934.82
				(5.41)	(3.68)	(41.61)	(42.15)	(58.46)	(57.67)
		NC=5	N(-1,1)	-5820.24	-6304.24	12408.89	12106.41	12460.50	13424.55
				(7.02)	(6.23)	(511.16)	(120.14)	(118.35)	(102.23)
			N(0,1)	-6309.82	-6698.06	12999.76	13079.13	13407.16	14323.92
				(12.03)	(7.16)	(75.65)	(175.49)	(96.01)	(63.99)
	1000	NC=3	N(-1,1)	-7319.89	-7692.63	16457.94	15815.14	16002.61	16752.82
				(4.50)	(4.00)	(757.09)	(316.39)	(159.94)	(157.72)
			N(0,1)	-7904.40	-8223.16	16893.86	16903.31	17231.09	17944.54
				(8.74)	(6.15)	(77.45)	(76.41)	(82.51)	(78.87)
		NC=5	N(-1,1)	-11492.11	-12408.24	24826.35	24164.35	24787.43	26724.80
				(12.10)	(4.65)	(760.98)	(332.38)	(157.20)	(188.19)
			N(0,1)	-12512.72	-13374.05	26098.78	26151.95	26851.18	28681.39
				(10.06)	(2.07)	(110.21)	(109.89)	(150.05)	(151.15)
n=20	500	NC=3	N(-1,1)	-7038.33	-7370.65	15213.21	14868.38	15170.13	15928.34
				(5.33)	(6.52)	(529.74)	(189.30)	(100.11)	(98.04)
			N(0,1)	-7683.27	-8100.82	15853.94	15866.16	16257.14	17093.66
				(10.77)	(10.49)	(76.68)	(76.48)	(102.68)	(79.29)
		NC=5	N(-1,1)	-11309.91	-12210.79	23143.15	23021.44	23724.76	25638.76
				(11.36)	(4.66)	(347.08)	(314.69)	(352.01)	(351.19)
			N(0,1)	-12305.96	-13221.44	25017.43	25082.75	25807.30	27747.64
				(13.10)	(6.61)	(137.05)	(124.77)	(154.44)	(84.94)
	1000	NC=3	N(-1,1)	-14375.09	-15133.84	33404.57	29718.81	30398.45	31839.40
				(7.77)	(7.46)	(7552.50)	(204.38)	(197.86)	(209.58)
			N(0,1)	-15238.40	-16105.85	31905.62	31931.76	32667.98	34384.95
				(6.29)	(6.40)	(100.40)	(97.75)	(125.08)	(148.13)
		NC=5	N(-1,1)	-22839.51	-24772.09	47519.03	46679.39	47958.95	51672.31
				(12.56)	(4.52)	(883.79)	(357.09)	(179.98)	(185.61)
			N(0,1)	-24916.32	-26791.09	50082.28	50216.08	51679.22	55533.91
				(10.55)	(8.51)	(244.88)	(252.99)	(293.65)	(229.47)

Table 5. Model Selection Frequencies

test leng.	samp. size	# of categ.	abil. distr.	true model	Selected by DIC				Selected by PsBF				Selected by AIC				Selected by BIC				Selected by LR			
					GR	GP	P	R	GR	GP	P	R	GR	GP	P	R	GR	GP	P	R	GR	GP	P	R
n=10	500	NC=3	N(-1,1)	GR	2	8	0	0	4	6	0	0	9	1	0	0	9	1	0	0	-	10	0	0
				GP	3	7	0	0	0	10	0	0	3	7	0	0	3	7	0	0	-	10	0	0
				P	0	1	9	0	0	1	9	0	1	0	9	0	0	0	10	0	-	1	9	0
				R	0	1	0	9	0	3	0	7	0	0	0	10	0	0	0	10	-	0	0	10
			N(0,1)	GR	3	7	0	0	9	1	0	0	9	1	0	0	9	1	0	0	-	10	0	0
				GP	1	9	0	0	0	10	0	0	1	9	0	0	1	9	0	0	-	10	0	0
				P	0	1	9	0	0	1	9	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	3	0	7	0	1	1	8	0	0	0	10	0	0	0	10	-	0	0	10
		NC=5	N(-1,1)	GR	9	1	0	0	10	0	0	0	6	4	0	0	6	4	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	1	9	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	1	9
			N(0,1)	GR	2	8	0	0	2	8	0	0	10	0	0	0	10	0	0	0	-	9	1	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	3	7	0	0	1	9	0	0	1	9	0	0	0	10	0	-	1	9	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
	1000	NC=3	N(-1,1)	GR	2	8	0	0	1	9	0	0	2	7	1	0	2	7	1	0	-	9	1	0
				GP	0	10	0	0	3	7	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	4	6	0	0	4	6	0	0	0	10	0	0	0	10	0	-	1	9	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
			N(0,1)	GR	8	2	0	0	8	2	0	0	10	0	0	0	10	0	0	0	-	10	0	0
				GP	0	10	0	0	6	4	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	2	8	0	0	2	8	0	0	0	10	0	0	0	10	0	-	1	9	0
				R	0	0	0	10	0	2	1	7	0	0	0	10	0	0	0	10	-	0	0	10
		NC=5	N(-1,1)	GR	10	0	0	0	10	0	0	0	4	6	0	0	4	6	0	0	-	9	1	0
				GP	1	9	0	0	0	10	0	0	1	9	0	0	1	9	0	0	-	9	1	0
				P	0	2	8	0	0	4	6	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
			N(0,1)	GR	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
n=20	500	NC=3	N(-1,1)	GR	4	6	0	0	6	4	0	0	5	4	1	0	5	4	1	0	-	9	1	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
			N(0,1)	GR	2	8	0	0	7	3	0	0	10	0	0	0	10	0	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	1	9	0	0	0	10	0	0	0	10	-	0	0	10
		NC=5	N(-1,1)	GR	8	2	0	0	10	0	0	0	7	3	0	0	7	3	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	-	2	8	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
			N(0,1)	GR	5	5	0	0	10	0	0	0	10	0	0	0	10	0	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	-	1	9	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
	1000	NC=3	N(-1,1)	GR	9	1	0	0	2	8	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	1	9	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	1	9	0	0	0	10	0	0	0	10	-	0	0	10
			N(0,1)	GR	9	1	0	0	4	6	0	0	10	0	0	0	10	0	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	2	8	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
		NC=5	N(-1,1)	GR	10	0	0	0	10	0	0	0	3	7	0	0	3	7	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	-	0	10	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10
			N(0,1)	GR	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	-	10	0	0
				GP	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	0	-	10	0	0
				P	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	0	-	1	9	0
				R	0	0	0	10	0	0	0	10	0	0	0	10	0	0	0	10	-	0	0	10

Appendix A: MATLAB Code Used for calculating PsBF

```
 cvlog.m
-------------------------------------------------------------------------
% Condition 2222GRM: data generated following GRM; n=20; N=1000; NC=5, theta~N(0,1)
n=20; N=1000; ncat=5; thm=0;
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:

% Cross-Validation Dataset
load gr2222v.txt; cvdat = gr2222v; cvloglik=zeros(10,4);

% CV log-likelihood of GRM
for z=1:10
    cur_GRM=z
    a=estgrm(1:n,z);
    % reading b1, b2, b3, and b4
    for s=1:n
        b1(s)=estgrm(n+(4*s-3),z);
        b2(s)=estgrm(n+(4*s-2),z);
        b3(s)=estgrm(n+(4*s-1),z);
        b4(s)=estgrm(n+(4*s),z);
    end
    cv=zeros(N,1);
    for j=1:N
        resp=zeros(1,n);
        resp=cvdat(j,:);
        ind_cv_grm
        cv(j)=cvj;
    end
    cvloglik(z,1)=sum(cv);
end

% CV log-likelihood of GPCM
for z=1:10
    cur_GPCM=z
    a=estgpcm(1:n,z); b=estgpcm(n+1:2*n,z);
    % reading tau2, tau3, tau4, and tau5
    for s=1:n
        tau2(s)=estgpcm(2*n+(4*s-3),z);
        tau3(s)=estgpcm(2*n+(4*s-2),z);
        tau4(s)=estgpcm(2*n+(4*s-1),z);
        tau5(s)=estgpcm(2*n+(4*s),z);
    end
    cv=zeros(N,1);
    for j=1:N
        resp=zeros(1,n);
        resp=cvdat(j,:);
        ind_cv_gpcm
        cv(j)=cvj;
    end
    cvloglik(z,2)=sum(cv);
end

% CV log-likelihood of PCM
for z=1:10
    cur_PCM=z
    a=ones(n,1);  b=estpcm(1:n,z);
    % reading tau2, tau3, tau4, and tau5
    for s=1:n
        tau2(s)=estpcm(n+(4*s-3),z);
```

```
        tau3(s)=estpcm(n+(4*s-2),z);
        tau4(s)=estpcm(n+(4*s-1),z);
        tau5(s)=estpcm(n+(4*s),z);
    end
    cv=zeros(N,1);
    for j=1:N
        resp=zeros(1,n);
        resp=cvdat(j,:);
        ind_cv_gpcm
        cv(j)=cvj;
    end
    cvloglik(z,3)=sum(cv);
end

% CV log-likelihood of RSM
for z=1:10
    cur_RSM=z
    a=ones(n,1);  b=estrsm(1:n,z);
    tau2=estrsm(n+1,z)*ones(n,1);
    tau3=estrsm(n+2,z)*ones(n,1);
    tau4=estrsm(n+3,z)*ones(n,1);
    tau5=estrsm(n+4,z)*ones(n,1);
    cv=zeros(N,1);
    for j=1:N
        resp=zeros(1,n);
        resp=cvdat(j,:);
        ind_cv_gpcm
        cv(j)=cvj;
    end
    cvloglik(z,4)=sum(cv);
end

% PsBF
dlmwrite('cvloglik.txt', cvloglik, '     ');
```

---

ind_cv_gpcm.m

```
% When the number of categories is 5
% 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,ncat); tt=zeros(K,n,ncat); 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)+tau2(i)));
        tt(t,i,3) = exp(a(i)*(k(t)-b(i)+tau2(i) + k(t)-b(i)+tau3(i)));
        tt(t,i,4) = exp(a(i)*(k(t)-b(i)+tau2(i) + k(t)-b(i)+tau3(i)+ k(t)-b(i)+tau4(i)));
        tt(t,i,5) = exp(a(i)*(k(t)-b(i)+tau2(i) + k(t)-b(i)+tau3(i)+ k(t)-b(i)+tau4(i)
                    + k(t)-b(i)+tau5(i)));
        denom(t,i) = 1 + tt(t,i,2) + tt(t,i,3) + tt(t,i,4) + tt(t,i,5);
    end
end for t=1:K
    for i=1:n
        for w=1:ncat
            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);
        elseif resp(i)==3
            lik=lik*pofc(t,i,3);
        elseif resp(i)==4
            lik=lik*pofc(t,i,4);
        else
            lik=lik*pofc(t,i,5);
        end
    end
    L(t)=lik;
end

% to cmpute a posterior probability of ability
for t=1:K
    prob(t)=L(t)*normpdf(k(t),thm,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
------------------------------------------------------------------------
% When the number of categories is 5
% 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,ncat); tt=zeros(K,n,ncat-1); 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) )));
        tt(t,i,3) = 1/(1 + exp(-a(i)*(k(t) - b3(i) )));
        tt(t,i,4) = 1/(1 + exp(-a(i)*(k(t) - b4(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) - tt(t,i,3);
        pofc(t,i,4) = tt(t,i,3) - tt(t,i,4);
        pofc(t,i,5) = tt(t,i,4);
    end
end

% to make it sure that the sum of pofc's(prob. of category) is the unity
for t=1:K
     for i=1:n
        totpofc(t,i)=pofc(t,i,1)+pofc(t,i,2)+pofc(t,i,3)+pofc(t,i,4)+pofc(t,i,5);
        pofc(t,i,1) = pofc(t,i,1) / totpofc(t,i);
        pofc(t,i,2) = pofc(t,i,2) / totpofc(t,i);
        pofc(t,i,3) = pofc(t,i,3) / totpofc(t,i);
        pofc(t,i,4) = pofc(t,i,4) / totpofc(t,i);
        pofc(t,i,5) = pofc(t,i,5) / totpofc(t,i);
    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);
        elseif resp(i)==3
            lik=lik*pofc(t,i,3);
        elseif resp(i)==4
            lik=lik*pofc(t,i,4);
        else
            lik=lik*pofc(t,i,5);
        end
    end
    L(t)=lik;
end

% to cmpute a posterior probability of ability
for t=1:K
    prob(t)=L(t)*normpdf(k(t),thm,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
------------------------------------------------------------------------
```

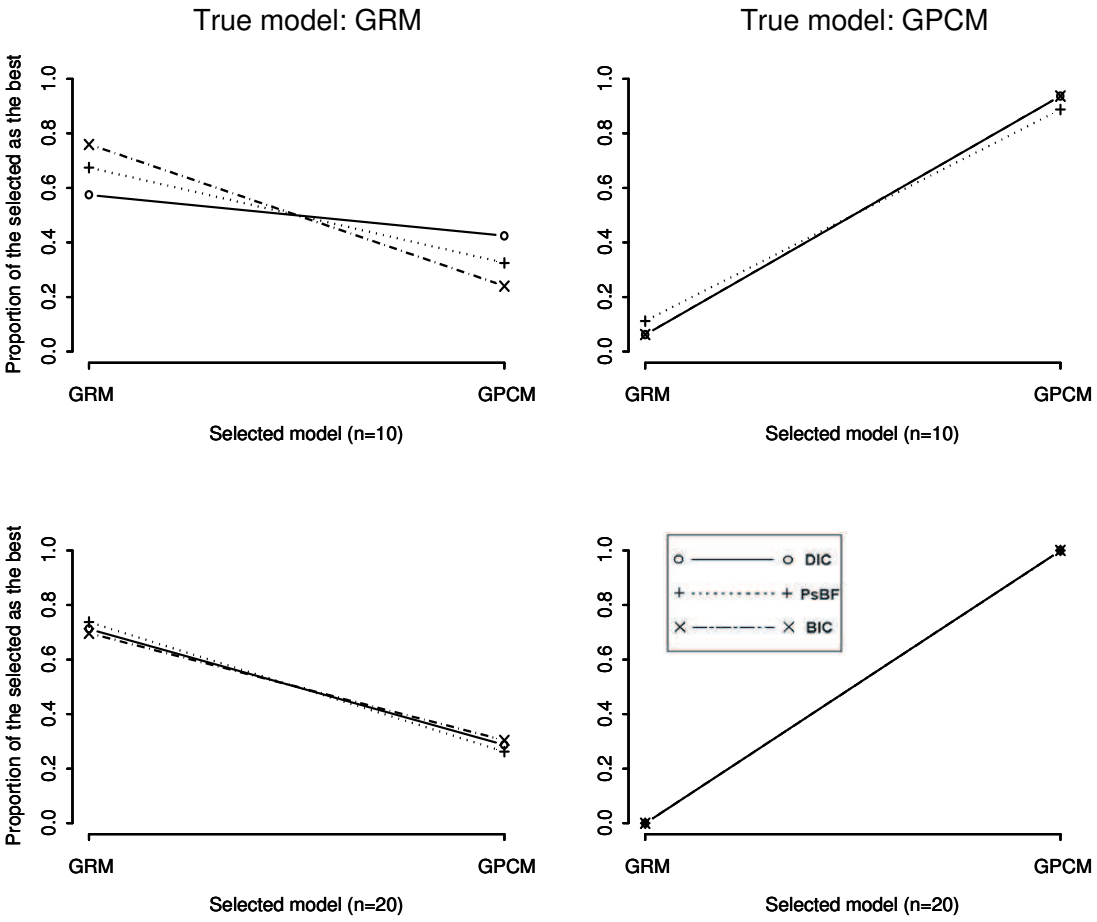

Figure 1: Model Selection Proportions by Test Length (DIC, PsBF and BIC)

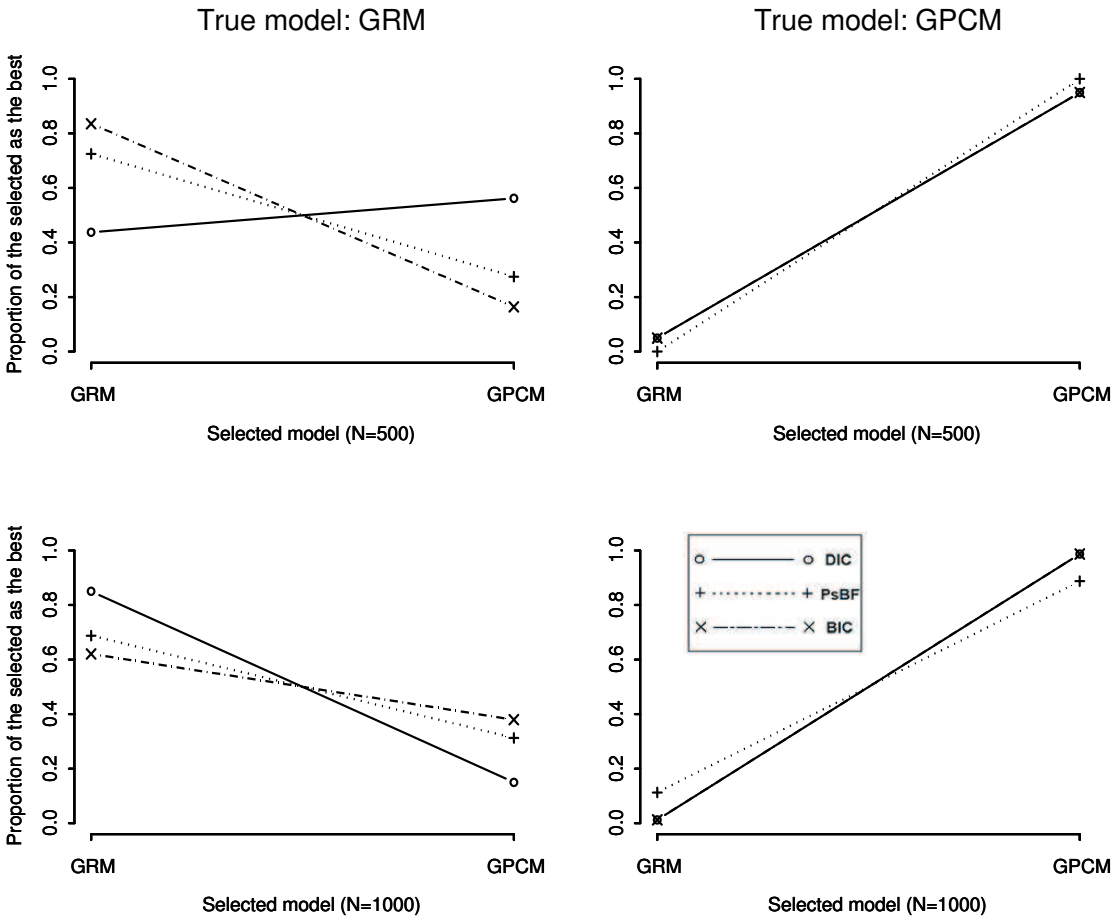

Figure 2: Model Selection Proportions by Sample Size (DIC, PsBF and BIC)

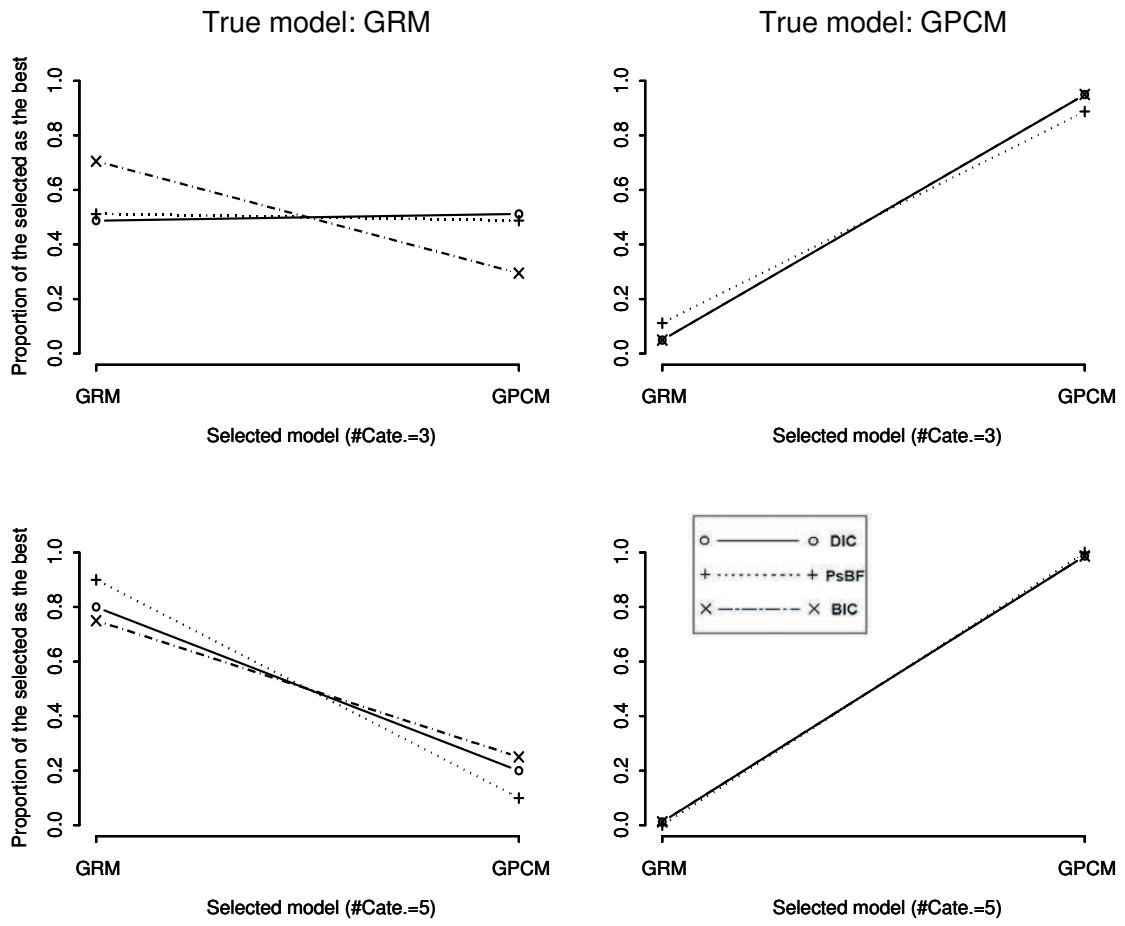

Figure 3: Model Selection Proportions by Number of Categories (DIC, PsBF and BIC)

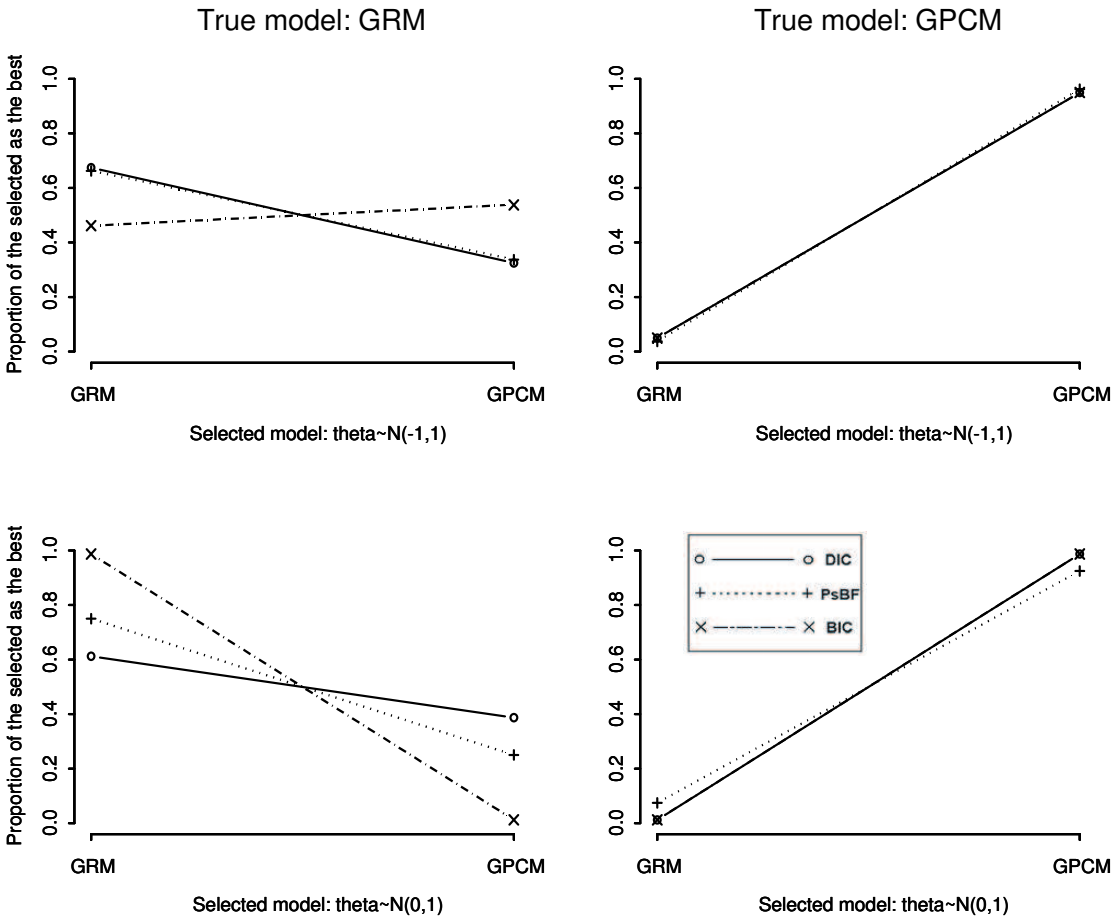

Figure 4: Model Selection Proportions by Ability Distribution (DIC, PsBF and BIC)