Applied Psychological Measurement
Volume 33 Number 7
October 2009 499-518
© 2009 SAGE Publications
10.1177/0146621608327800
http://apm.sagepub.com
hosted at
http://online.sagepub.com

Model Selection Indices for Polytomous Items

Taehoon Kang
Allan S. Cohen
Hyun-Jung Sung

This study examines the utility of four indices for use in model selection with nested and nonnested polytomous item response theory (IRT) models: a cross-validation index and three information-based indices. Four commonly used polytomous IRT models are considered: the graded response model, the generalized partial credit model, the partial credit model, and the rating scale model. In a simulation study, comparisons among the four indices suggest that model selection is dependent to some extent on the particular conditions simulated. Overall, the Bayesian information criterion index appears to be most accurate in selecting the correct polytomous IRT model. Results are presented from analysis of a real data set to illustrate the use of the four indices for selecting an appropriate model.

Keywords: *item response theory; model selection; AIC; BIC; DIC; cross-validation log likelihood*

Tests composed of constructed-response items are often scored in more than two categories and as a result need to be fit with a polytomous model. Under item response theory (IRT), determining which polytomous model is the best fit to the data is a difficult task owing in part to model complexity. Selection of the wrong model may have serious consequences such as those that have been reported: IRT test equating (Bolt, 1999; Camilli, Wang, & Fesq, 1995; Dorans & Kingston, 1985; Kaskowitz & De Ayala, 2001), parameter estimation (DeMars, 2005; Wainer & Thissen, 1987; Walker & Beretvas, 2003; Zenisky, Hambleton, & Sireci, 1999), computer adaptive testing (Ackerman, 1991; De Ayala, Dodd, & Koch, 1992; Greaud-Folk & Green, 1989), person–fit assessment (Drasgow, 1982; Meijer & Sijtsma, 1995), and analysis of differential item functioning (Bolt, 2002). Even so, relatively little research has been conducted regarding the accuracy of different methodologies for determining the selection of an appropriate polytomous IRT model given the data. With respect to polytomous models, this has had the unfortunate consequence of many researchers simply selecting a model with which they are familiar or for which software is available (Embretson & Reise, 2000).

In this regard, Maydeu-Olivares, Drasgow, and Mead (1994) used the ideal observer index (IOI) to compare fit to the data of the graded response model (GRM; Samejima,

Authors' Note: From the University of California, Los Angeles (TK); University of Georgia, Athens (ASC); and Pearson, Tulsa, Oklahoma (HJS). Please address correspondence to Taehoon Kang, National Center for Research on Evaluation, Standards, & Student Testing, 300 E. Young Drive North, GSE&IS Bldg, Suite 320, Los Angeles, CA 90095 e-mail: taehoonkang@gmail.com.

1969) and the generalized partial credit model (GPCM; Muraki, 1992), two commonly used polytomous IRT models. Maydeu-Olivares et al. concluded that either model would be equally appropriate in most practical applications. If this result is correct, then it would suggest that it may well be a matter of indifference which of the two models is used. Given that the two models do not imply the same psychological ordering among score categories, this may possibly not be an appropriate solution. The GRM uses a proportional odds model in which all response categories for an item are collapsed into two categories for estimating the boundary characteristic curves. In this way, a series of two-parameter models are used for model calibration. Under the GPCM, however, the focus is on the relative difficulty of each step needed to transition from one category to the next in an item because the GPCM uses an adjacent odds model.

When marginal maximum likelihood estimates of model parameters are available, previous research has suggested that information-based statistics such as Akaike's information criterion (AIC; Akaike, 1974) or Schwarz's Bayesian information criterion (BIC; Schwarz, 1978) may be useful for model selection for dichotomous items (Kang & Cohen, 2007). Although significance tests are not possible with these statistics, they do provide estimates of the relative differences between models. When Bayesian methods are used to obtain estimates of parameters for polytomous models, however, little evidence has been presented as to which model selection criteria are more useful, although several indices are available. One is cross-validation log likelihoods (CVLL; Bolt, Cohen, & Wollack, 2001; Geisser & Eddy, 1979; Gelfand & Dey, 1994; Kang & Cohen, 2007) and another is the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002). Regardless of whether IRT models are nested, the four indices (AIC, BIC, DIC, and CVLL) all can be used for model selection. There is no reported research however that has studied this issue with respect to polytomous items under either marginal maximum likelihood estimation (MMLE) or Bayesian estimation. In this study, comparative information was examined regarding the accuracy of each of these indices for determining the best-fitting model to the data for marginal maximum likelihood and full-Bayesian estimation.

To provide a basis for making comparisons among the four statistics, model selection was examined for the following four polytomous IRT models: the rating scale model (RSM; Andrich, 1978), the partial credit model (PCM; Masters, 1982), the GPCM, and the GRM. The first three models, the RSM, PCM, and GPCM, are hierarchically related to each other, although they are not nested with respect to the GRM. The GRM is included because it is not hierarchically related and it has the same number of parameters as the GPCM.

In the sequel, the four polytomous IRT models and the four model selection indices are described followed by two studies. A simulation study was conducted to explore the relative behaviors of the four indices for the four polytomous IRT models under practical testing conditions. An empirical example is then presented to illustrate use of the four model selection indices for selecting the most appropriate polytomous IRT model.

Polytomous IRT Models

This study deals with the four commonly used polytomous IRT models noted above: the RSM, PCM, GPCM, and GRM. The first three models, the RSM, PCM, and GPCM, represent an extension of the two-parameter logistic model (2PLM) and are described by Thissen and Steinberg (1986) as divide-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. By convention, $\tau_{0i}=0$, $\sum_{k=1}^{m}\tau_{ki}=0$, and $\exp\sum_{k=0}^{0}\alpha_i[\theta_j-(\beta_i-\tau_k)]=1$ in equation (1) for identification. If the α_i are fixed at 1 across items, equation (1) reduces to the PCM. If in addition τ values are the same for each category across items, equation (1) further reduces to the RSM. Consequently, the RSM, PCM, and GPCM are nested models.

The GRM, however, is not a divide-by-total model. Instead, Thissen and Steinberg (1986) refer to the GRM as a difference model. It also can be viewed as a generalization of the 2PLM that in the logistic form uses the two-parameter logistic function to model boundary characteristic curves. These curves represent the probability of a response higher than a given category x. It is convenient in the GRM model to convert the $x=0,\ldots,m$ category scores into $x=1,\ldots,m+1$ categories. If P^*_{ijx} is used to denote the boundary probability for examinee j to receive a category score larger than x on item i, then the boundary curve is given by

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

The difference between adjacent categories is used to determine the probability of a particular item score. Thus, in the GRM, the probability that examinee j receives a category score x on 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 $P^*_{ij(m+1)}=0$.

The GRM is distinguished from the GPCM and its two nested models by the fact that the GRM requires a two-step process to compute the conditional probability for an examinee responding in a particular category. Myung, Pitt, Zhang, and Balasubramanian (2001) note that model complexity is a function of the number of free parameters of a model and its functional form. For example, $y=ax$ and $y=x^a$ both have the same number of parameters but they function differently given the data. With respect to the GRM and GPCM, both models have the same number of parameters for fitting each item but it is not necessarily the case that they have the same model complexity, as the way in which each treats the data is functionally different.

Model Selection Indices

A desirable model selection process should consider a model's complexity as well as its goodness of fit (Akaike, 1974; Forster, 1999). A good model should fit not only the current data set but also could explain another data set obtained through the same process. With an unnecessarily complicated model (i.e., one that Sober, 2002, notes violates the principle of parsimony and can actually reduce predictive accuracy) predictions about future data sets actually can worsen.

The DIC (Spiegelhalter et al., 2002) is composed of a Bayesian measure of fit or adequacy called the posterior mean deviance, $\overline{D}$, and a penalty function for model complexity, p_D, based on the number of free parameters in the model:

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

where $\overline{D(\theta)}$, the posterior mean of the deviances, 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 DIC was developed by Spiegelhalter et al. to find the best-fitting model without rewarding overparameterization. The model with the smallest DIC is assumed to be the one that would best predict a replicate data set of the same structure. In the IRT context, calculation of the CVLL is done as follows: First, two samples are drawn, a calibration sample, Y_{cal}, in which the examinees are randomly sampled from the whole data, and a cross-validation sample, Y_{cv}, which is a second sample that is randomly drawn from the remaining examinees. The calibration sample is used to update prior distributions of model parameters to posterior distributions. The likelihood of the Y_{cv} for a model can be computed using the updated posterior distribution as a prior (Bolt & Lall, 2003):

$$P(Y_{cv}|\text{Model}) = \int P(Y_{cv}|\theta, Y_{cal}, \text{Model}) f_\theta(\theta|Y_{cal}, \text{Model}) d\theta, \tag{5}$$

where $P(Y_{cv}|\theta, Y_{cal}, \text{Model})$ represents the conditional likelihood and $f_\theta(\theta|Y_{cal}, \text{Model})$ the conditional posterior distribution. An estimate of CVLL for a model is obtained as the logarithm of $P(Y_{cv}|\text{Model})$ in equation (5). The quality of competing models can be checked by cross-validation. For the CVLL, the predictive accuracy of each model can be evaluated in terms of log likelihoods using cross-validation samples from the same population. The preferred model can be determined through a direct comparison of individual CVLLs. When more than two models are compared, the decision rule is that the model with the largest CVLL is the best (Bolt et al., 2001; Spiegelhalter, Thomas, Best, & Gilks, 1996).

Sahu (2002) used the CVLL and the DIC for model selection with dichotomous IRT models. Although the simulation conditions used by Sahu were somewhat limited, results suggested that both the CVLL and DIC might have some utility for IRT model selection. Kang and Cohen (2007) likewise found that these two indices had potential utility for selecting dichotomous IRT models.

Information-based indices such as the AIC and BIC have been found to be useful because they tend to strike a balance between improvement in model fit as a result of increasing the number of model parameters and parsimony (De Boeck, Wilson, & Acton, 2005). The AIC has two components representing goodness of fit and complexity. In this study, the first component, the deviance, is defined as –2 × log(marginal maximum likelihood). The second component is $2 \times p$, where p is the number of estimated parameters. This second component can be interpreted as a penalty function for overparameterization. A criticism of the AIC is that it is not asymptotically consistent because sample size is not directly involved in its calculation (Ostini & Nering, 2005; Schwarz, 1978; Sclove, 1987). The AIC tends to select the more saturated of the models in very large samples (Forster, 2004; Janssen & De Boeck, 1999).

The BIC achieves asymptotic consistency by penalizing overparameterization with the use of a logarithmic function of the sample size in which p is multiplied by a number proportional to N. As a result, BIC tends to select the simpler of the models relative to the AIC, when the sample size is large. Results from these two statistics do not always agree (Lin & Dayton, 1997; Lubke & Muthén, 2005) because of the different penalty functions.

With respect to the performance of AIC and BIC, it is not necessarily clear which of these provides the better results in specific cases (Wagner & Timmer, 2001) nor is it certain that these two indices are appropriate for comparing nonnested models with different types of parameters or scales (Hong & Preston, 2005; Ostini & Nering, 2005). Even though there is relatively common use of these indices, their performance in IRT applications has not been widely examined. In the case of dichotomous IRT models, Kang and Cohen (2007) found that AIC and BIC were useful for selecting the correct one-parameter and two-parameter models but both were poor at choosing the correct three-parameter model. No such evidence has been presented with respect to selection of polytomous IRT models.

Simulation Study Comparing Model Selection Indices

In this simulation study, the behavior of the four indices is explored using simulated data. It was designed to systematically evaluate each of the four indices on the four polytomous IRT models under known generating models and parameters. Data were generated with four different polytomous IRT models under a variety of practical testing conditions. The intent of this study is to be able to inform the use of each of these model selection indices under conditions normally encountered in practical testing situations.

Method

Simulation design. Data were simulated for each of the four polytomous IRT models described above (RSM, PCM, GPCM, and GRM), two test lengths ($n = 10$ and 20), two sample sizes ($N = 500$ and 1,000), and two numbers of categories per item (NC = 3 and 5).

Computer program codes written by the authors with the MATLAB software (MathWorks, Natick, Massachusetts, 2001) were used to generate the data sets for each simulation condition. (The MATLAB code is available from the authors on request.)

The two test lengths were intended to simulate broad-range achievement tests having either moderate or large numbers of polytomously scored items. The means of the location parameters were drawn from across a range of –1.5 to +1.5 on the ability metric therefore to simulate a broad-range achievement test. This resulted in a set of simulated item location parameters that were spread across the intended range of the test. Discrimination parameters for the GPCM and GRM were randomly sampled from a lognormal (0, 0.5^2) distribution. For five-category items, four location parameters of each item were randomly drawn from normal distributions with standard deviations of 1.0 and means of –1.5, –0.5, 0.5, and 1.5, respectively. The resulting difficulties sometimes needed to be adjusted to meet the ordering assumption for the GRM. This only occurred 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 ($b_{1i}, \ldots, b_{4i}$) for an item was used as the item difficulty parameter (b_i) and the difference between b_i and the b_{ki}s were taken as the category parameters, τ_{ki}s. θ values were randomly drawn from a normal (0,1) distribution.

For an item with three categories, the generating values for the location parameters were obtained as the mean of each pair of adjacent generating parameters for the respective five-category item. That is, the mean of b_{1i} and b_{2i} was taken as the new b_{1i} and the mean of b_{3i} and b_{4i} was taken as the new b_{2i} for items simulated to have three categories. For the GPCM, the process of deriving item difficulty parameters and the category parameters was the same as for five-category items.

The generating parameters for each of the items are shown in Tables 1 and 2 for the three-category and five-category items, respectively. 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 × 2 test lengths × 2 sample sizes × 2 numbers of categories. Fifty replications were generated for each condition. For each generated data set, parameters for the same four polytomous models were estimated using both MMLE and Markov chain Monte Carlo (MCMC) algorithms implemented in the computer softwares PARSCALE 4.1 (Muraki & Bock, 2003) and WinBUGS 1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2003), respectively. To evaluate the performance of the four model selection indices, comparisons were made for the proportions of times each index selected the correct model.

Parameter estimation. Marginal maximum likelihood estimates of item parameters were obtained using the computer program PARSCALE. The PARSCALE program provides an estimate of $-2 \times \log$(marginal maximum likelihood) for each set of items calibrated. AIC and BIC were estimated using parameter estimates obtained from

Table 1
Generating Item Parameters (Number of Categories = 3)

	GRM			GPCM		
Item	*a*	b_1	b_2	*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
M	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

Note: $\tau_0 = 0$ for model idenfitication. GRM = graded response model; GPCM = generalized partial credit model.

PARSCALE. A Gibbs sampling algorithm was used as implemented in the computer program WinBUGS to estimate model parameters under MCMC or full-Bayesian solution. (Examples of the PARSCALE and WinBUGS codes are available on request from the authors.)

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 MCMC sampler is run until the iterations can be assumed to be sampled from stationary distributions. The initial set of iterations are known as burn-in iterations and discarded. Subsequent iterations are retained for use in estimating model parameters. The estimate is sampled from the posterior after each iteration subsequent to burn-in and the MCMC estimate is taken as the mean over these iterations.

WinBUGS software includes a utility for estimation of DIC for each model. The CVLL was calculated with the posterior item parameter estimates for a calibration sample and examinees' responses in a cross-validation sample. To derive the posterior distributions for each parameter sampled, it is first necessary to specify prior distributions. In this study,

Table 2
Generating Item Parameters (Number of Categories = 5)

	GRM					GPCM				
Item	a	b_1	b_2	b_3	b_4	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
M	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

Note: $\tau_0 = 0$ for model idenfitication. GRM = graded response model; GPCM = generalized partial credit model.

the following priors were used for items with $m + 1$ categories for the GPCM: θ_j: normal (0, 1), $(j = 1, \ldots, N), a_i$ lognormal (0, 1),$(i = 1, \ldots, T), b_i$ normal(0, 1),τ_{ki}: normal(0, 10), $(k = 1, \ldots, m - 1)$, 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_{ki} indicates the location of category k relative to item i's difficulty. In addition, the following constraints were used in the GPCM: $\sum_{k=0}^{m} \tau_{ki} = 0$, and $\tau_{0i} = 0$ in equation (1). For the GRM, the following priors were used: θ_j: normal(0, 1), $(j = 1, \ldots, N), a_i$: lognormal (0, 1), $(i = 1, \ldots, T)$, b_{1i}: normal(0, 10), b_{ki}: normal(0, 10) $I(b_{(k-1)i})$, $(k = 2, \ldots, m)$, where the notation $I(b_{(k-1)i},)$ indicates that b_{ki} is sampled to be larger than $b_{(k-1)i}$.

Checking convergence. Determination of a suitable burn-in for each of the models 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 that can be used to determine an appropriate length for the burn-in. The Gelman-Rubin convergence statistic (*R*; Brooks & Gelman, 1998) was calculated along chains for each model. For the RSM, PCM, and GPCM, *R* converged to 1 for all model parameters after less than 500 iterations. The GRM, however, required about 4,000 iterations to reach convergence.

When the history graph of the Markov chain was drawn for each item parameter using three different initial values, the GPCM converged relatively quickly whereas the GRM required substantially more iterations. On the basis of these results, a conservative estimate of 5,000 iterations for the burn-in was used for all four models in this study. For each chain, an additional 6,000 iterations were run after discarding the burn-in iterations. Estimates of model parameters were based on the means of the sampled values from these 6,000 iterations.

Results

Recovery of item parameters. The quality of the recovery of model parameters was checked on all the 50 data sets for each of the simulated conditions. Parameter recovery was evaluated using both the root mean square error (RMSE) and product moment correlation (r) 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).

The recovery results for all parameters in the four polytomous IRT models were good ($\bar{r} \geq .93$) for both MMLE and MCMC calibrations. The recovery results, the averages of 50 RMSEs and correlations for all four models, appeared to be similar. So those for only the GPCM and GRM parameters are reported in Table 3 (and those for the PCM and RSM results are not provided). The average RMSEs for both models tend to be smaller as the sample size or test length increases. The results in Table 3 are in general agreement with recovery results reported in the literature (e.g., DeMars, 2003; Kim & Cohen, 2002).

Behavior of model selection indices. The frequencies of model selections for the four different indices (DIC, CVLL, AIC, and BIC) are shown in Figures 1 to 3. In these graphs, marginal results are plotted for the three factors of test length, sample size, and number of categories, respectively.

In Figure 1, the model selection frequencies are plotted as separate histograms for each of the four indices for the different test lengths ($n = 10$, and $n = 20$). The frequencies were calculated marginally so that 200 data sets (50 data sets × 2 sample sizes × 2 number of categories) are summarized in each plot. When the true (i.e., the generating) model was the GPCM, PCM or RSM, the four indices performed relatively well in selecting the correct (i.e., the generating) model. A clear increase in correct model selection was evident for the longer test ($n = 20$). Regardless of test length, however, when the true model was the GRM, the GPCM tended to be selected by the DIC and CVLL in roughly a third of the data sets. The AIC and BIC showed very good performances regardless of true model.

Figure 2 presents 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 rather than the GRM as the better model about 2/3 of the time, and when $N = 1{,}000$, DIC performed somewhat better, selecting the correct model, the GRM, approximately 86% ($= 171/200$) of the time. The CVLL did not show any improvement for the larger sample size (actually becoming a bit worse) when the true model was the GRM or GPCM. When the true model was the GPCM, PCM, or RSM, the four indices performed well for both sample sizes.

Table 3
Recovery Statistics: Average RMSEs (Average *r*s) Between Generating and Estimated Item Parameters

Test	Sample	No. of	GPCM (MMLE)					GRM (MMLE)				
Length (n)	Size (N)	Categories	a	b	τ_1	τ_2	τ_3	a	b_1	b_2	b_3	b_4
10	500	3	0.14 (0.98)	0.10 (0.98)	0.17 (0.95)			0.16 (0.98)	0.23 (0.98)	0.14 (0.98)		
		5	0.12 (0.99)	0.08 (0.98)	0.27 (0.98)	0.21 (0.97)	0.22 (0.97)	0.13 (0.98)	0.29 (0.98)	0.14 (0.99)	0.10 (0.99)	0.17 (0.98)
	1,000	3	0.10 (0.99)	0.06 (0.99)	0.11 (0.97)			0.11 (0.98)	0.14 (0.99)	0.10 (0.99)		
		5	0.09 (0.99)	0.05 (0.99)	0.18 (0.99)	0.15 (0.98)	0.15 (0.99)	0.09 (0.99)	0.19 (0.99)	0.10 (0.99)	0.08 (0.99)	0.12 (0.99)
20	500	3	0.12 (0.97)	0.09 (0.97)	0.12 (0.94)			0.15 (0.96)	0.18 (0.99)	0.17 (0.97)		
		5	0.11 (0.98)	0.07 (0.98)	0.21 (0.98)	0.19 (0.96)	0.19 (0.98)	0.12 (0.97)	0.25 (0.98)	0.12 (0.99)	0.12 (0.99)	0.19 (0.97)
	1,000	3	0.09 (0.99)	0.06 (0.99)	0.09 (0.97)			0.11 (0.98)	0.12 (0.99)	0.12 (0.99)		
		5	0.08 (0.99)	0.05 (0.99)	0.15 (0.99)	0.12 (0.99)	0.14 (0.99)	0.08 (0.99)	0.16 (0.99)	0.08 (1.00)	0.08 (0.99)	0.12 (0.98)
			GPCM (MCMC)					GRM (MCMC)				
			a	b	τ_1	τ_2	τ_3	a	b_1	b_2	b_3	b_4
10	500	3	0.15 (0.98)	0.10 (0.97)	0.18 (0.93)			0.16 (0.96)	0.25 (0.98)	0.20 (0.98)		
		5	0.12 (0.98)	0.08 (0.98)	0.26 (0.96)	0.21 (0.97)	0.21 (0.96)	0.13 (0.98)	0.33 (0.97)	0.18 (0.99)	0.16 (0.99)	0.24 (0.97)
	1,000	3	0.10 (0.99)	0.07 (0.99)	0.11 (0.97)			0.11 (0.98)	0.17 (0.99)	0.16 (0.99)		
		5	0.09	0.05 (0.99)	0.18 (0.98)	0.16 (0.99)	0.15 (0.98)	0.09 (0.99)	0.22 (0.99)	0.14 (0.99)	0.13 (0.99)	0.18 (0.99)
20	500	3	0.12 (0.97)	0.09 (0.97)	0.12 (0.94)			0.15 (0.97)	0.19 (0.98)	0.18 (0.97)		
		5	0.11 (0.98)	0.07 (0.98)	0.22 (0.97)	0.19 (0.98)	0.19 (0.97)	0.13 (0.97)	0.30 (0.97)	0.13 (0.99)	0.12 (0.99)	0.20 (0.97)
	1,000	3	0.09 (0.99)	0.06 (0.99)	0.09 (0.97)			0.11 (0.98)	0.13 (0.99)	0.12 (0.99)		
		5	0.08 (0.99)	0.05 (0.99)	0.15 (0.99)	0.13 (0.99)	0.14 (0.99)	0.08 (0.99)	0.16 (0.99)	0.09 (1.00)	0.08 (0.99)	0.12 (0.99)

Note: RMSE = root mean square error; GPCM = generalized partial credit model; GRM = graded response model; MMLE = marginal maximum likelihood estimation; MCMC = Markov chain Monte Carlo.

Figure 1
Model Selection Frequencies by Test Length

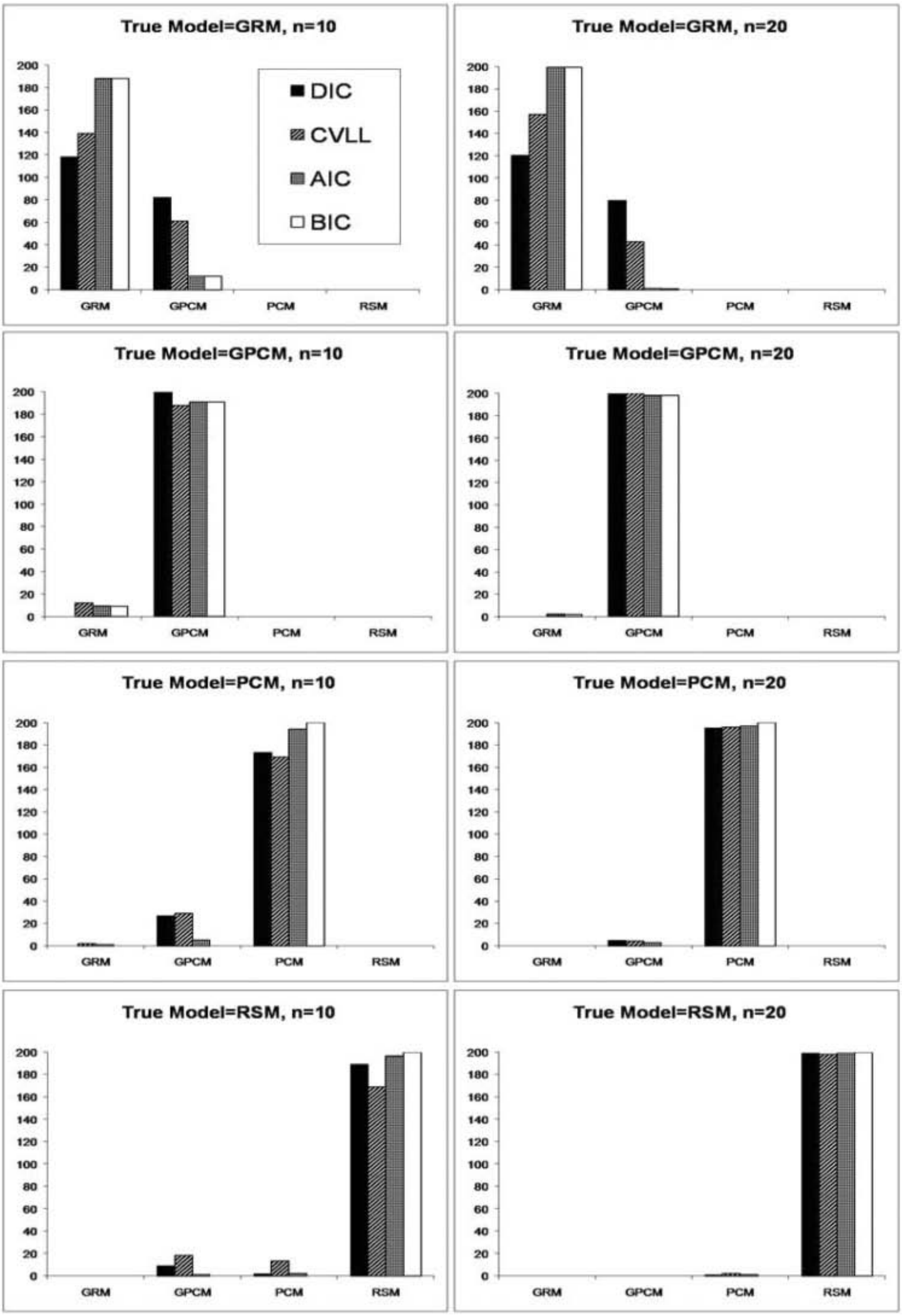


Figure 2
Model Selection Frequencies by Sample Size

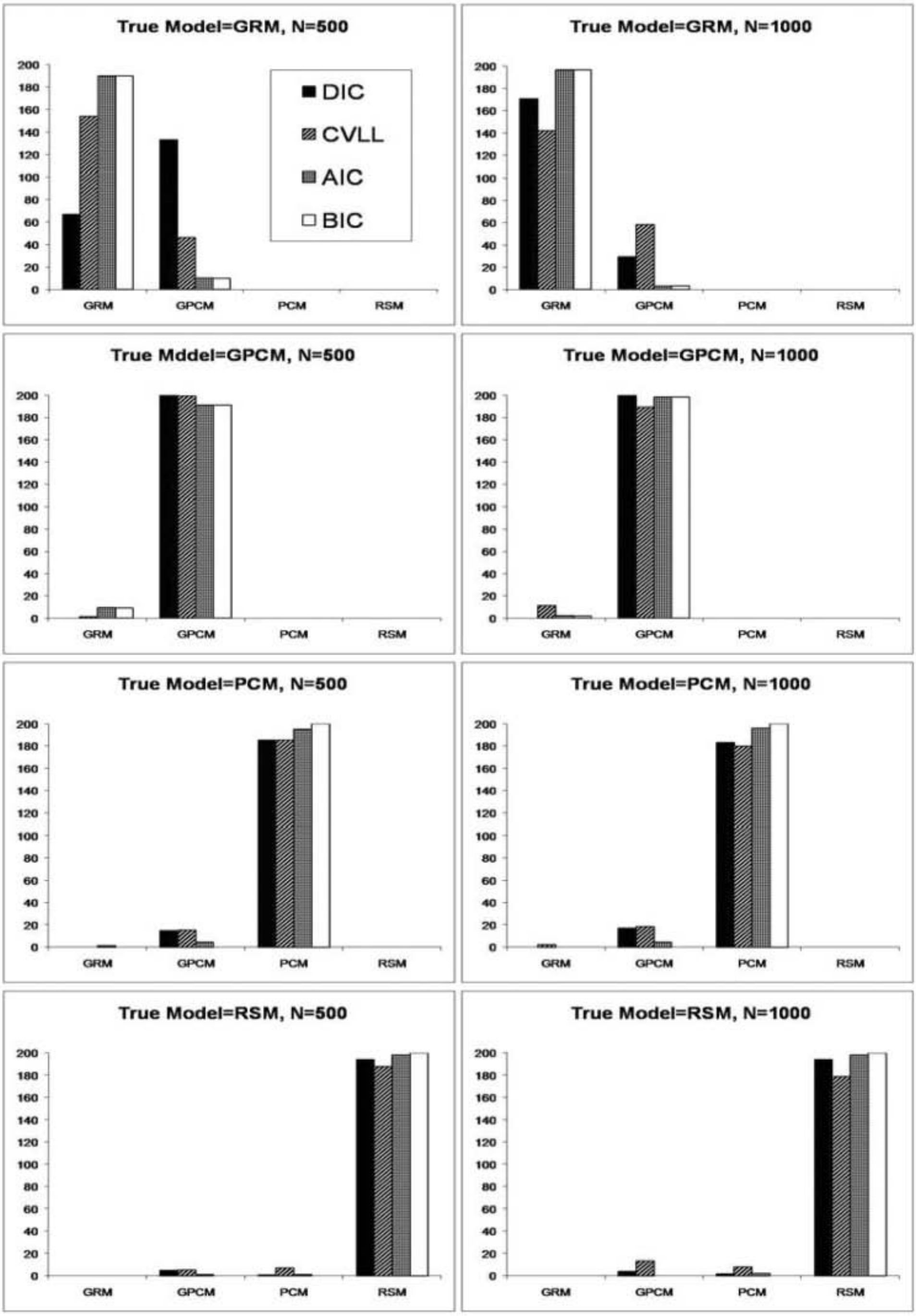
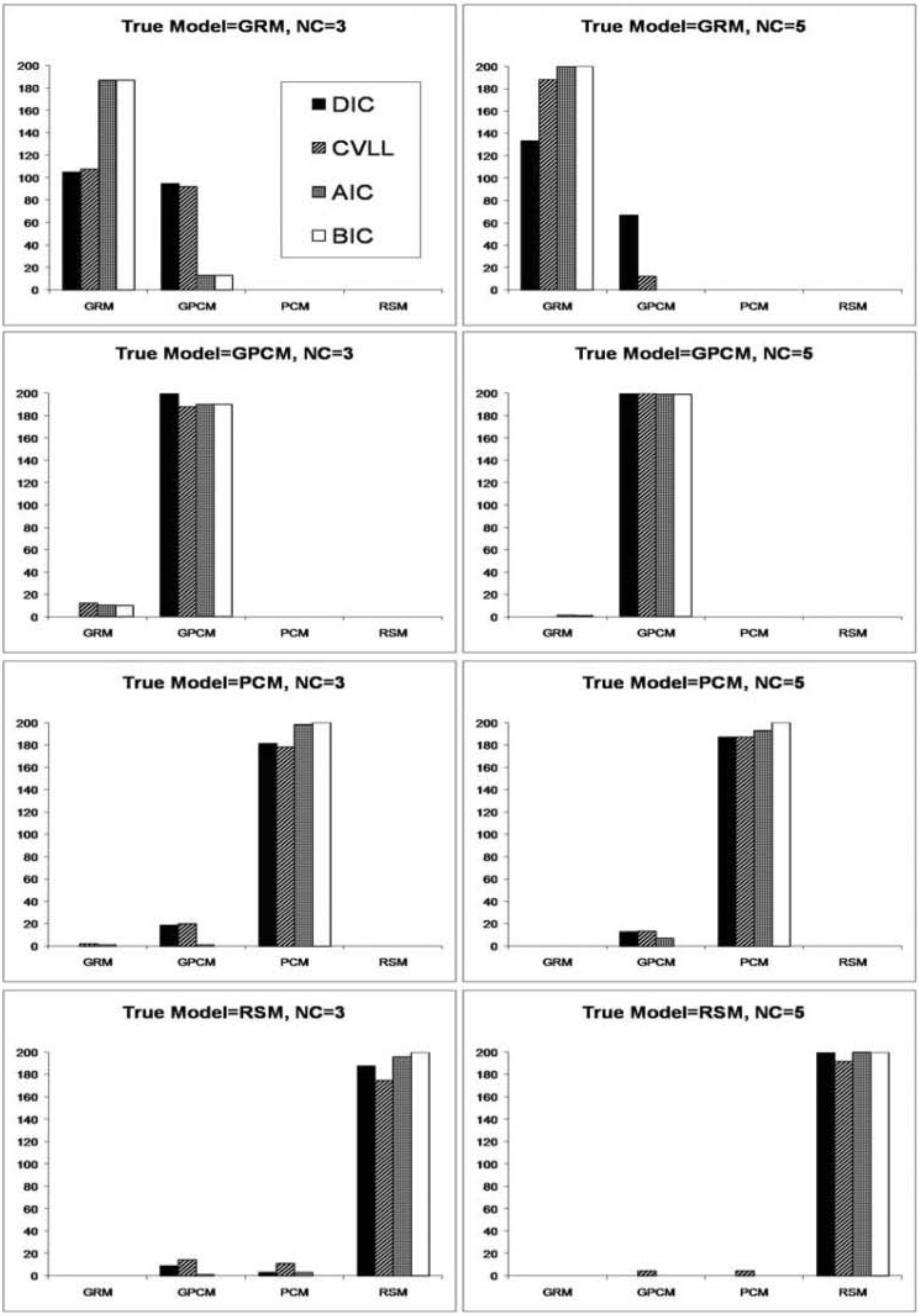


Figure 3
Model Selection Frequencies by Number of Categories



Overall, the AIC and BIC could find the generating model very accurately irrespective of sample size. Especially, the BIC was correct in 100% of the data sets when the correct model was the PCM or RSM.

In Figure 3, the performance of the model selection indices appeared to be sensitive to the number of categories. For a test with five-category items, the DIC, CVLL, AIC, and BIC selected the true GRM with 67%, 94%, 100%, and 100% accuracy, respectively, compared with 53%, 54%, 94%, and 94% accuracy for tests with three-category items. When the true model was the GPCM or RSM, however, all four indices worked almost perfectly in selecting the correct model for the tests with five-category items. The performance of the DIC and CVLL showed a slight improvement in finding the true PCM for the larger number of categories.

From Figures 1 to 3, first of all, it is evident that in the larger sample size ($N = 1{,}000$) and the five-category conditions, all four indices appeared to select the correct model most of the time. If the true model was one of the three nested models (i.e., the RSM, PCM, or GPCM), all indices likewise performed well, particularly for the longer test length ($n = 20$). For the GRM as the correct model, however, DIC and CVLL showed good performances only when the number of categories of items was five and sample size was 1,000. DIC in particular performed poorly at selecting the GRM for the smaller sample size ($N = 500$). In brief, all four model selection indices worked well, when the data were generated with the RSM, PCM, or GPCM. The DIC and CVLL, however, did not perform as well at selecting the GRM rather than the GPCM, when the data were generated with the GRM. When the sample was small ($n = 10$, $N = 500$, $\text{NC} = 3$), both AIC and BIC were less accurate (i.e., 82%) in locating the correct GRM. The percentage of correct selections of each of the four indices over all 1,600 data sets ($1{,}600 = 4$ IRT models $\times$ 2 test lengths $\times$ 2 sample sizes $\times$ 2 numbers of categories $\times$ 50 data sets) was 87% (1,394) for the DIC, 89% (1,416) for the CVLL, 98% (1,563) for the AIC, and 98% (1,574) for the BIC. When only the three nested models, RSM, PCM, and GPCM, were considered, the accuracy levels increased to 96% for the DIC, 93% for the CVLL, 98% for the AIC, and 99% for the BIC.

Table 4 shows the average differences of the model selection indices (generating model minus calibrating model), when the true model was the GRM. Similar results were obtained when the true models were the GPCM, PCM, and RSM, but those results are not reported in this article. Because the differences relating DIC, AIC, and BIC were negative and the differences in CVLLs were positive in most cases, the findings through these average difference (Δ) values appeared to be consistent with a tendency found in Figures 1 to 3. As shown in Table 4, the GRM and the GPCM were often very close to each other: The absolute values of the Δs between the two models were smaller than the other differences for every index in all conditions. This is not surprising, as these models use the same number of parameters to fit an item.

Example Study of IRT Model Selection

A real data study is presented to illustrate the use and relative performance of the four indices. The example presented below is to choose the best polytomous IRT model for an English placement test data set.

Table 4
Average Differences (*SD*) of Model Selection Indices (Generating Model – Calibrating Model) When Data Were Generated With the GRM

Test Length (n)	Sample Size (N)	No. of Categories	Δ^{DIC}_{GR-GP}	Δ^{DIC}_{GR-P}	Δ^{DIC}_{GR-R}	Δ^{CVLL}_{GR-GP}	Δ^{CVLL}_{GR-P}	Δ^{CVLL}_{GR-R}
10	500	3	1.03 (5.25)	–219.60 (34.80)	–573.47 (47.17)	1.92 (8.03)	112.29 (10.60)	262.78 (13.63)
		5	5.26 (12.46)	–353.65 (50.33)	–1243.32 (73.57)	15.07 (14.57)	174.23 (12.99)	588.08 (20.00)
	1,000	3	–5.05 (7.72)	–452.15 (60.44)	–1168.63 (68.85)	–25.13 (39.91)	254.09 (41.70)	593.11 (40.54)
		5	–39.97 (20.55)	–751.01 (87.43)	–2536.72 (128.27)	41.67 (10.41)	395.32 (16.59)	1298.99 (31.61)
20	500	3	6.72 (17.76)	–425.08 (52.09)	–1257.79 (81.60)	0.17 (9.80)	229.81 (18.55)	667.93 (27.22)
		5	3.84 (21.04)	–710.93 (56.89)	–2551.69 (100.12)	38.78 (10.91)	377.51 (17.06)	1340.34 (15.02)
	1000	3	–8.74 (12.89)	–875.37 (66.25)	–2549.75 (114.11)	–4.24 (21.87)	384.54 (23.88)	1306.08 (35.86)
		5	–78.04 (27.81)	–1508.18 (117.61)	–5274.61 (155.78)	81.05 (5.95)	827.54 (23.09)	2728.22 (36.80)
			Δ^{AIC}_{GR-GP}	Δ^{AIC}_{GR-P}	Δ^{AIC}_{GR-R}	Δ^{BIC}_{GR-GP}	Δ^{BIC}_{GR-P}	Δ^{BIC}_{GR-R}
10	500	3	–4.01 (3.82)	–146.62 (23.62)	–494.89 (40.53)	–4.01 (3.82)	–104.48 (23.62)	–414.82 (40.53)
		5	–36.85 (82.96)	–375.91 (40.75)	–1260.85 (69.75)	–36.85 (82.96)	–333.76 (40.75)	–1104.91 (69.75)
	1,000	3	–9.00 (5.06)	–305.09 (39.93)	–1010.51 (54.12)	–9.00 (5.07)	–256.01 (39.93)	–917.26 (54.12)
		5	–54.57 (14.87)	–758.95 (67.17)	–2537.42 (117.43)	–54.57 (14.87)	–709.88 (67.17)	–2355.83 (117.43)
20	500	3	–12.17 (6.67)	–357.47 (38.14)	–1178.81 (71.36)	–12.17 (6.67)	–273.18 (38.13)	–1014.44 (71.36)
		5	–65.39 (17.50)	–756.94 (47.88)	–2591.94 (99.20)	–65.39 (17.50)	–672.65 (47.88)	–2267.41 (99.20)
	1,000	3	–23.13 (10.26)	–726.03 (54.07)	–2378.02 (106.85)	–23.13 (10.25)	–627.88 (54.07)	–2186.62 (106.85)
		5	–129.05 (23.94)	–1521.49 (96.80)	–5280.84 (141.61)	–129.05 (23.94)	–1423.33 (96.80)	–4902.94 (141.61)

Note: GRM = graded response model; AIC = Akaike’s information criterion; BIC = Bayesian information criterion; DIC = deviance information criterion; CVLL = cross-validation log likelihoods.

Table 5
Model Selection Indices for Five Polytomous Items From 2005 English Placement Test Data

	Model Selection Methods			
Model	DIC	CVLL	AIC	BIC
RSM	40207.80	–18856	42675.51	42729.57
PCM	40170.60	–18846	42542.78	42692.94
GPCM	40195.80	–18875	42529.34	42709.53
GRM	40270.70	–18872	42564.47	42744.66

Note: RSM = rating scale model; PCM = partial credit model; GPCM = generalized partial credit model; GRM = graded response model; DIC = deviance information criterion; CVLL = cross-validation log likelihoods; AIC = Akaike's information criterion; BIC = Bayesian information criterion.

Description of the English Test Data

Data for this study were taken from responses of 22,102 college freshmen to an English placement test used at a Midwestern university in 2005. The test was used to place incoming freshmen into courses in the introductory composition sequence. The reading comprehension section of this test included five passages, each with five items. In this study, each passage was scored as a single polytomous model with scores ranging from 0 to 5, with possible scores ranging from 0 to 25.

Item parameters for the four IRT models were all calibrated on a random sample (without replacement) of 3,000 examinees using both MMLE and MCMC methods. Then each of the four model selection indices was calculated for each of the IRT models. A second random sample of 3,000 examinees was drawn to obtain the CVLL estimates. The mean over all five polytomous items in the calibration sample was 17.55 ($SD = 4.68$) and 17.71 ($SD = 4.62$) in the cross-validation sample. The first eigenvalue from a principal components analysis of the polychoric correlation matrix for the five items was 2.76, indicating 55.24% of the total variance was attributed to the first factor. All other eigenvalues were less than 0.63, suggesting that the items measured a unidimensional construct. And Cronbach's α was estimated as .80.

Results for the Example Study

Except for AIC, model selection results were consistent in choosing the best polytomous IRT model among the indices (see Table 5). The DIC, CVLL, and BIC all indicated that the PCM could be selected as the best model, but the AIC indicated the GPCM was the best fit. The GRM was ranked as the 3rd or 4th best fit to the data by all four indices.

As mentioned earlier, the PCM is more parsimonious than the GPCM, requiring fewer parameters. The selection of the GPCM by AIC was consistent with previous research indicating AIC tends to select the more complicated model as sample size increases (Forster, 2004; Janssen & De Boeck, 1999). The product–moment correlations between item and total test scores ranged from .68 to .76, indicating each of the items had high discriminations. Based on these results, the PCM, one of the family of Rasch models, appeared to be the best fit.

The simulation study presented above considered two sample sizes, 500 and 1,000, and the performances of AIC and BIC were quite similar in almost every condition. The empirical example, however, suggests that sample size may have an important effect on the behavior of these indices and could usefully be studied further.

Conclusions and Discussions

This study investigated the performance of four model selection indices for use in selecting the most appropriate polytomous IRT model among four candidate models. In the 32 simulated conditions, some indices appeared to function better under some conditions and better for some models. These results agree with previous research (Kang, 2006; Kang & Cohen, 2007; Li, Cohen, Kim, & Cho, 2006; Sung & Kang, 2006). BIC generally appeared to be the most accurate and consistent in selecting the true (i.e., the generating) model, although the performance of the other three indices was also good in many cases. Model selection with AIC was almost the same as with BIC except for a few conditions in which either the PCM or RSM were the generating models.

When the true model was the GRM, however, the performances of DIC and CVLL were less accurate in some conditions, as shown in Figures 1 to 3. In addition, Δ values of DIC and CVLL between the GRM and GPCM were much smaller for data generated under the GRM. For conditions in this study with the larger data sets (i.e., 1,000 examinees, and five-category items), however, the performance of these model selection indices with respect to the GRM was consistently very good. This would imply that a large data set is required to select the correct GRM for the DIC and CVLL. In contrast, in most conditions, both AIC and BIC had low misidentification rates when the generating model was either the GPCM or GRM. This would suggest that these two indices would be useful, even in a relatively small data set. In this study, two different estimation methods were used to calculate model selection indices: MCMC for DIC and CVLL, and MMLE for AIC and BIC. In the GRM calibration, the average RMSE values of MCMC were consistently equal to or larger than those of MMLE, as shown in Table 3. The relatively poor GRM–MCMC calibration seems to be a possible explanation of the poor performance of DIC and CVLL in finding the true GRM.

In this study, it was assumed that there exists a generating or true model among those in the set of candidate models. This is not a necessary assumption. It is appropriate to consider a situation in which the true model may not be among those in the candidate set. This is in fact what may be faced in a practical testing situation, as it is not possible to know which model is the true model. Consequently, the best that can be done with real data is to examine the fit for available models and carefully consider the results given a theoretical rationale for each model and the purpose for which the test was developed. Assuming a set of theoretically appropriate models is assembled, however, one criterion for selecting the best model should be how well predictive accuracy is retained for different data sets (Forster, 1999; Hitchcock & Sober, 2004; Sober, 2002). The selection is based on the model with the smallest AIC, BIC, DIC, or negative CVLL value. These measures provide an indication of the relative distances from the unknown true model. Although it cannot be said for certain that the true model has been selected with real data, these indices are useful as long as the task is to inform the selection of the most appropriate model among the set of reasonable candidate models.

References

Ackerman, T. A. (1991).The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*, 13-24.

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

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

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education, 12*, 383-406.

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.

Brooks, S. P., & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434-455.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79-96.

De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education, 5*, 17-34.

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

DeMars, C. E. (2003). Sample size and the recovery of nominal model item parameters. *Applied Psychological Measurement, 27*, 275-288.

DeMars, C. E. (2005, August). *Scoring subscales using multidimensional item response theory models*. Poster presented at the annual meeting of the American Psychology Association. Retrieved March 6, 2009, from http://www.jmu.edu/assessment/wmlibrary/subscaledemo.doc

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*, 249-262.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6*, 297-308.

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

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*. Unpublished manuscript, University of Wisconsin–Madison. Retrieved August 15, 2007, from 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.

Greaud-Folk, V., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373-390.

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

Hong, H., & Preston, B. (2005). *Nonnested model selection criteria*. Unpublished manuscript, Department of Economics, Columbia University, New York. Retrieved September 1, 2006, from http://www.columbia.edu/bp2121/nonnestmsc.pdf

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

Kang, T. (2006). *Model selection methods for unidimensional and multidimensional IRT models*. Unpublished doctoral dissertation, University of Wisconsin–Madison.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*, 331-358.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25*, 39-52.

Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*, 25-41.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2006, April). *Model selection methods for mixture dichotomous IRT models*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco

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.

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.

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

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18*, 245-256.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*, 261-272.

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

Muraki, E., & Bock, R. D. (2003). *PARSCALE (version 4.1): IRT item analysis and test scoring for rating-scale data* [Computer software]. Chicago: Scientific Software.

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

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

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.

Sober, E. (2002). Instrumentalism, parsimony, and the Akaike framework. *Philosophy of Science, 69*, 112-123.

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*, 583-640.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. (1996). *BUGS 0.5* Bayesian Inference Using Gibbs Sampling Manual (version ii).* Cambridge, UK: MRC Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS User Manual (version 1.4)*. Retrieved March 6, 2009, from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf

Sung, H.-J., & Kang, T. (2006, April). *Choosing a polytomous IRT model with Bayesian model selection methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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

Wagner, M., & Timmer, J. (2001). Model selection in non-nested hidden Markov models for ion channel gating. *Journal of Theoretical Biology, 208*, 439-450.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*, 339-368.

Walker, C. M. & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*, 255-275.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (1999). *Effects of item local dependence on the validity of IRT item, test, and ability statistics*. Washington, DC: Association of American Medical Colleges. Retrieved February 1, 2008, from http://www.aamc.org/students/mcat/research/monograph5.pdf