Bayesian Multidimensional IRT Models With a Hierarchical Structure
Yanyan Sheng and Christopher K. Wikle
Educational and Psychological Measurement 2008; 68; 413 originally published online Dec 3, 2007;
DOI: 10.1177/0013164407308512

The online version of this article can be found at:
http://epm.sagepub.com/cgi/content/abstract/68/3/413

Published by:
SAGE Publications
http://www.sagepublications.com

Additional services and information for Educational and Psychological Measurement can be found at:

Email Alerts: http://epm.sagepub.com/cgi/alerts

Subscriptions: http://epm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations (this article cites 20 articles hosted on the SAGE Journals Online and HighWire Press platforms):
http://epm.sagepub.com/cgi/content/refs/68/3/413
Bayesian Multidimensional IRT Models With a Hierarchical Structure

Yanyan Sheng

Southern Illinois University Carbondale

Christopher K. Wikle

University of Missouri–Columbia

As item response models gain increased popularity in large-scale educational and measurement testing situations, many studies have been conducted on the development and applications of unidimensional and multidimensional models. Recently, attention has been paid to IRT-based models with an overall ability dimension underlying several ability dimensions specific for individual test items, where the focus is mainly on models with dichotomous latent traits. The purpose of this study is to propose such models with continuous latent traits under the Bayesian framework. The proposed models are further compared with the conventional IRT models using Bayesian model choice techniques. The results from simulation studies as well as actual data suggest that (a) such models can be developed; (b) compared with the unidimensional IRT model, the proposed models better describe the actual data; and (c) the use of the proposed IRT models and the multidimensional model should be based on different beliefs about the underlying dimensional structure of a test.

Keywords: item response theory; hierarchical MIRT model; general and specific abilities; unidimensional model; multidimensional model; MCMC; Gibbs sampling; Bayesian model choice

Much research has been conducted on the development and application of Bayesian unidimensional IRT models, where one ability dimension is assumed (e.g., Albert, 1992; Baker, 1998; Johnson & Albert, 1999; Patz & Junker, 1999), and multidimensional IRT (MIRT) models, where multiple ability dimensions are involved in one test (e.g., Béguin & Glas, 2001; Hoijtink & Molenaar, 1997; Lee, 1995). With the availability of powerful computers and the Markov chain Monte Carlo (MCMC) simulation techniques (e.g., Chib & Greenberg, 1995), the more complicated IRT models that incorporate an overall ability dimension underlying several

Authors’ Note: Please address correspondence to Yanyan Sheng, Department of Educational Psychology and Special Education, Wham 223, Mail Code 4618, Southern Illinois University, Carbondale, IL 62901; e-mail: ysheng@siu.edu.
ability dimensions specific for individual test items can be developed. Recent attempts have been made to develop such models with dichotomous latent traits (e.g., de la Torre & Douglas, 2004). These models are appropriate for cognitive diagnosis. But for more general purposes or applications such as ability estimation or equating, models with continuous latent traits are desired. In this article, we focus on developing such IRT models in the Bayesian framework.

One of the major considerations in measurement theory is to ensure that meaningful inferences are made from test scores, which requires fit of the empirical test data obtained from test indicators to a theoretical framework. This forms the basis for legitimizing any application of modern IRT models. However, it is well accepted among measurement theorists that none of the theoretical models fully represents a complex reality (van der Linden & Hambleton, 1997). Instead, the models are only a simplified approximation of the real world. Therefore, when employing IRT in a testing situation, one has to choose a model that provides the most complete description of the data.

The unidimensional and the multidimensional IRT models make different assumptions about the latent dimensions. However, in many applications, because of the lack of reliable statistical tests for dimensionality (e.g., De Champlain & Gessaroli, 1998; Gessaroli & De Champlain, 1996; Hattie, 1985; Hattie, Krakowski, Rogers, & Swaminathan, 1996), it is not easy to choose one model over the other. For instance, suppose an English test consists of three subtests: listening, reading, and writing. One can assume that all items are measuring a unified English ability dimension and fit a unidimensional model. On the other hand, if the test items are designed to measure specifically listening, reading, or writing, then a multidimensional model or, specifically, a multiunidimensional model (a special type of the multidimensional model, which is applicable under the situation when the overall test consists of unidimensional subtests; Sheng & Wikle, 2007) might be more appropriate. The key difference between the two models lies in whether a single composite score is sufficient or if subscores have to be reported for illustrating the examinee’s proficiency in answering all test items.

Consequently, the critical difference between the unidimensional model and the multiunidimensional model makes model comparison necessary as to which model leads to a more accurate estimate of the trait. In certain circumstances, one may want to report both composite and subtest scores. To this end, one may (a) implement a multiunidimensional model to obtain the subtest scores and then average them to get the composite score or (b) implement both the unidimensional and the multiunidimensional models. However, averaging may give rise to biased composite scores, and two separate implementations overlook the relationships between the latent abilities and could be time consuming. It is then reasonable to include overall ability as well as specific ability dimensions in the model, so that both composite and subtest scores can be obtained with one single implementation.
The purpose of the study is to propose such an IRT-based model under the Bayesian framework that incorporates an overall ability dimension underlying all test items and several ability dimensions specific for each subtest. This latent structure is comparable to that of the second-order factor model (Schmid & Leiman, 1957) in the factor analytic framework, and what is essentially novel about this study is the incorporation of such a latent structure into the IRT framework, so that the proposed model, while assuming a hierarchical structure for the underlying ability dimensions, focuses on the interaction between persons and items at the item level. An MCMC algorithm known as Gibbs sampling is used to implement the model, so that both item characteristics and person traits can be estimated simultaneously. To illustrate the Gibbs sampling procedure for the proposed model, a subset of *College Basic Academic Subjects Examination (CBASE) English* subject data is used.

The proposed model is more complicated than the unidimensional IRT model and thus should provide better fit to the actual test data. Model comparisons are thus carried out using Bayes factors (BFs) and Bayesian deviance information criteria (DICs; Spiegelhalter, Best, & Carlin, 1998). Because the adequacy of a model is directly related to model assumptions, when comparing a unidimensional model with the proposed model or a multiunidimensional model, one is automatically testing whether the unidimensionality assumption holds for actual test data. The Bayesian model comparison techniques provide an alternative method of checking model assumptions.

The significance of this study lies in the fact that a new IRT model can offer several advantages to measurement validation that are not realized in existing models. Specifically, the advantages of the current approach are threefold. First, the proposed model incorporates one general ability for the overall test and several specific abilities for the subtests. Thus, both levels of the ability dimensions can be estimated with one implementation, so that both composite scores and subscores are reported. Second, the fully Bayesian approach allows one to estimate item parameters and examinee abilities at the same time while incorporating uncertainty of the item estimates in calculations of uncertainty about abilities for examinees (e.g., Albert, 1992; Béguin & Glas, 2001; Patz & Junker, 1999). And last, comparable to the $\chi^2$ test in factor analysis, the Bayesian model comparisons are confirmatory in nature, which has been lacking in the IRT literature (Segall, 2002).

The remainder of the article is organized as follows. The second section, “IRT Models,” reviews the unidimensional and multiunidimensional IRT models, and the next section describes the proposed hierarchical MIRT models in the Bayesian framework. This is followed by a section that illustrates the Bayesian model selection techniques. To evaluate model performances, a simulation study is conducted to assess the proposed IRT models regarding their parameter recovery under different test situations, and the results are summarized in the section “Simulation Study 1.” The next section discusses another simulation study using Bayesian model choice techniques to compare the proposed hierarchical MIRT models with the
unidimensional and multiunidimensional IRT models. Then, an example is provided in the section “A Real Data Example,” where the proposed models are implemented on a subset of CBASE English subject data. Finally, we close with a few remarks in the section “Discussion and Conclusions.”

IRT Models

In this article, we focus primarily on the two-parameter (2P) normal ogive (probit) models.

Unidimensional IRT Model

The unidimensional IRT model provides the simplest framework in modeling the person–item interaction by assuming one ability dimension. Suppose a test consists of \( k \) multiple choice items, each measuring a single unified ability, \( \theta \). Define \( y_{ij} \) as

\[
y_{ij} = \begin{cases} 
1, & \text{if person } i \text{ answers item } j \text{ correctly,} \\
0, & \text{if person } i \text{ answers item } j \text{ incorrectly,}
\end{cases} \quad i = 1, \ldots, n, \quad j = 1, \ldots, k.
\]

Then, the probability of person \( i \) obtaining a correct response for item \( j \) can be defined as follows:

\[
P(y_{ij} = 1|\theta_i, \alpha_j, \gamma_j) = \Phi(\alpha_j \theta_i - \gamma_j) = \int_{-\infty}^{\alpha_j \theta_i - \gamma_j} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \quad (1)
\]

where \( \alpha_j \) is a scalar parameter describing the item discrimination, \( \gamma_j \) is associated with item difficulty \( \beta_j \) such that \( \gamma_j = \alpha_j \beta_j \), and \( \theta_i \) is a scalar ability parameter.

Multiunidimensional IRT Model

One of the multivariate extensions of the unidimensional model is the multiunidimensional IRT model, which allows separate inferences to be made about an examinee for each specific ability dimension being measured by a test question. Consider a \( K \)-item test consisting of \( m \) subtests, each containing \( k_v \) multiple choice items that measure one ability dimension. With a probit link, the probability of person \( i \) obtaining the correct response for item \( j \) of the \( v \)th subtest can be defined as follows:

\[
P(y_{vij} = 1|\theta_{vi}, \alpha_v, \gamma_v) = \Phi(\alpha_v \theta_{vi} - \gamma_v) = \int_{-\infty}^{\alpha_v \theta_{vi} - \gamma_v} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \quad (2)
\]
where $\alpha_{ij}$ and $\theta_{vi}$ are scalar parameters representing the item discrimination and the examinee ability in the $v$th ability dimension and $\gamma_{ij}$ is a scalar parameter indicating the location in that dimension where the item provides maximum information. For a more detailed illustration of the model under the fully Bayesian framework, see Lee (1995) or Sheng and Wikle (2007).

The Proposed Bayesian Hierarchical MIRT Models

Suppose a $K$-item test consists of $m$ subtests, each containing $k_v$ multiple choice items, where $v = 1, 2, \ldots, m$. We let $y_{vij}$ denote the binary response from the $i$th examinee to the $j$th item of the $v$th subtest, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k_v$. Then, the proposed 2P hierarchical MIRT model is defined to have the following probability function:

$$P(y_{vij} = 1) = \Phi(\alpha_{ij}\theta_{vi} - \gamma_{ij}),$$

(3)

where $\alpha_{ij}$ and $\gamma_{ij}$ are item parameters and $\theta_{vi}$ is the specific ability parameter corresponding to the $v$th subtest. One may note that this probability function has exactly the same form as that of the multidimensional model presented in Equation 2. However, incorporating the general ability dimension, the former specifies a slightly more complicated underlying latent structure. Two hierarchical MIRT models are proposed in this study based on different beliefs concerning the relationship between the general and specific abilities.

**Model 1.** First, we assume that each specific ability is a linear function of the general ability such that

$$\theta_{vi} = \beta_v \theta_{0i} + \epsilon_{vi},$$

(4)

where $\epsilon_{vi} \sim N(0, 1)$, $\theta_{0i}$ is the $i$th examinee’s ability parameter for the overall test, and $\beta_v$ is a measure of association between the general ability and the $v$th specific ability parameters. Combining Equations 3 and 4, the model probability function becomes

$$P(y_{vij}=1) = \Phi(\alpha_{ij}\beta_v \theta_{0i} - \gamma_{ij} + \alpha_{ij}\epsilon_{vi}).$$

One may note the similarity of this formulation with that of Bradlow, Wainer, and Wang’s (1999) testlet model, whose systematic component takes the form $\alpha_j \theta_{0i} - \gamma_j = \gamma_j$. It can be shown that the testlet model is a special case of Model 1 where $\beta_v = 1$. That is, if expressed in our context of specific and general abilities, each specific ability can differ from the general ability only by a constant (the error). Hence, by introducing an additional parameter $\beta_v$, Model 1 is appropriate
when the specific ability is any linear function of the general ability. Moreover, this proposed model allows one to specify a different distribution for \( \alpha \) or \( \gamma \) for each subtest, whereas the testlet model does not. The latter is hence limited in situations where, for instance, it is believed that items in a particular subtest are supposed to have very different characteristics from those in other subtests. This further illustrates that Model 1 is more general and thus offers more flexibility than the testlet model.

For this model, it is assumed that \( \theta_{0i} \sim N(0,1) \), \( \alpha_{ij} > 0 \), and \( p(\gamma_{ij}) \propto 1 \). A uniform prior can be assumed for \( \beta_i \). However, a sensitivity analysis indicates a conjugate \( N(1,1) \) prior does not result in very different parameter estimates, which suggests that the model is not sensitive to the choice of prior distributions for this parameter. In effect, the differences are only to the second decimals. Consequently, with the more informative normal prior for \( \beta_i \), the posterior estimates for item and person parameters can be obtained using the Gibbs sampling procedure.

**Model 2.** Alternatively, it can be assumed that the general ability is a linear combination of all the specific abilities so that

\[
\theta_{0i} = \sum_v \lambda_v \theta_{vi} + \varepsilon_i, \tag{5}
\]

where \( \varepsilon_i \sim N(0,1) \) and \( \lambda_v \) is the standardized regression coefficient for the \( v \)th specific ability parameter when both \( \theta_{0i} \) and \( \theta_{vi} \) are standard normal variates. Although both proposed MIRT models specify linear relationships between the general ability and the specific abilities, they differ in the nature of the association, which, as reflected in Equations 4 and 5, is comparable to the difference between factor analysis and principal components. This point has to be emphasized because it reflects the key difference between the two models. In particular, Model 1 assumes that the general ability is a factor for each specific ability, so that \( \theta_{vi} \sim N(\beta_v \theta_{0i}, 1) \), whereas Model 2 assumes that the general ability is a linear function of all the specific abilities and \( \theta_{0i} \sim N\left(\sum_v \lambda_v \theta_{vi}, 1\right) \).

The specific abilities are assumed to have a multivariate normal prior, \( \theta_i \sim N_m(\mathbf{0}, \mathbf{R}) \), where \( \mathbf{R} \) is a correlation matrix with 1s on the diagonal and correlation \( \rho_{st} \) between \( \theta_s \) and \( \theta_t \), \( s \neq t \) on the off diagonals. With this constraint imposed on \( \mathbf{R} \), difficulty arises in choosing its prior distribution. To solve the problem, we follow Lee’s (1995) solution by introducing an unconstrained covariance matrix \( \Sigma \), where \( \Sigma = [\sigma_{ij}]_{m \times m} \), so that the correlation matrix \( \mathbf{R} \) can be easily transformed from \( \Sigma \) using \( \rho_{st} = \sigma_{st} / \sqrt{\sigma_{ss} \sigma_{tt}} \) \( (s \neq t) \). An inverse Wishart distribution is assumed for \( \Sigma \) so that \( \Sigma \sim W^{-1}(I, m) \), where \( I \) is an \( m \times m \) identity matrix.

Similar to the specifications for Model 1, a conjugate \( N(1,1) \) normal prior is assumed for \( \lambda_v \), as a sensitivity analysis indicates that the results do not differ much in parameter estimation from those using a uniform prior for \( \lambda_v \). Furthermore, with
flat priors for $\gamma_{ij}$ and $\alpha_{ij}$, that is, $p(\gamma_{ij}) \propto 1$ and $\alpha_{ij} > 0$, the posterior estimates for item and person parameters can be obtained.

**Bayesian Model Choice Techniques**

In the Bayesian framework, the adequacy of the fit of a given model is evaluated using several model choice techniques, among which BF$s$ and Bayesian deviance are considered in this study. It should be noted that both measures provide model comparison criteria. Hence, they evaluate the fit of a model in a relative, not absolute, sense.

**Bayes Factor**

When a set of $s$ different Bayesian hierarchical models $M_1, \ldots, M_s$ are considered, the BF for comparing two models $M_i$ and $M_j$ is defined as $BF = p(y|M_i)/p(y|M_j)$, where $p(y|M) = \int p(y|\theta)p(\theta|M)d\theta$ is the marginal probability of the data $y$ (also referred to as the prior mean of the likelihood) with $\theta$ denoting all model parameters, and $p(\theta|M)$ is the prior density for the unknown parameters under the specific model $M$. This is the Bayesian analog of the likelihood ratio between two models, and describes the evidence provided by the data in favor of $M_i$ over $M_j$. BF$s$ allow comparison of nonnested models and ensure consistent results for model comparisons but are usually difficult to calculate because of the difficulty in exact analytic evaluation of the marginal density of the data. Some approximation methods, such as Laplace integration, the Schwarz criterion, and reversible jump, and so on have been proposed and developed (see Kass & Raftery, 1995, for a detailed description). In more complex modeling situations, MCMC provides another approximation for the marginal density. Although it is potentially unstable, previous studies show that it often produces results that are accurate enough for interpreting the BF$s$ (e.g., Carlin & Chib, 1993), and therefore, we use it in this study.

To estimate the marginal density, one can draw MCMC samples of the parameters, $\xi^{(1)}, \ldots, \xi^{(G)}$ and $\theta^{(1)}, \ldots, \theta^{(G)}$, so that $p(y|M)$ is approximated as $\left\{ (1/G) \sum_{g=1}^{G} L(y|\theta^{(g)}, \xi^{(g)})^{-1} \right\}^{-1}$, the harmonic mean of the likelihood values (Newton & Raftery, 1994).

**Bayesian Deviance**

The Bayesian DIC was introduced by Spiegelhalter et al. (1998) who generalized the classical information criteria to one that is based on the posterior distribution of
the deviance. This criterion is defined as $DIC = \tilde{D} + p_D$, where $\tilde{D} \equiv E_{\theta|y}(D) = E(−2\log L(y|\theta))$ is the posterior expectation of the deviance (with $L$ being the likelihood function, i.e., $L(y|\theta) = \prod p(\theta_i)(1−p(\theta_i))^{1−y}$, where $\theta$ denotes all model parameters, and $p(\theta)$ is the appropriate probability function for the model) and $p_D = E_{\theta|y}(D) - D(E_{\theta|y}(\theta)) = D - D(\bar{\theta})$ is the effective number of parameters (Carlin & Louis, 2000). Furthermore, $D(\bar{\theta}) = −2\log (L(y|\bar{\theta}))$, where $\bar{\theta}$ is the posterior mean. To compute the Bayesian DIC, MCMC samples of the parameters, $\xi^{(1)}, \ldots, \xi^{(G)}$ and $\theta^{(1)}, \ldots, \theta^{(G)}$, can be drawn with the Gibbs procedure; then $\tilde{D}$ could be approximated as $\tilde{D} = (1/G)\{-2\log \prod_{g=1}^{G} L(y|\theta^{(g)}, \xi^{(g)})\}$. Generally, more complicated models tend to provide better fits. Hence, penalizing for number of parameters makes the DIC a more reasonable measure to use. However, unlike the BF, the DIC is not invariant to parameterization and sometimes can produce unrealistic results.

**Simulation Study 1**

The two hierarchical MIRT models are based on different beliefs concerning the relationship between the general ability and specific abilities. That is, Model 1 assumes that each specific ability is a linear function of the general ability, $\theta_{vi} = \beta_v \theta_0i + \xi_{vi}$, whereas Model 2 assumes that the general ability is a linear combination of the specific abilities, $\theta_0i = \sum \lambda_v \theta_{vi} + \xi_i$. In Model 2, the correlations between the specific ability dimensions are reflected through the specification of the prior, $\theta_i \sim N(0, R)$. On the other hand, the correlations between the specific abilities are realized through their relationships with the general ability in Model 1. To compare the two hierarchical models, their performance in item parameter recovery was evaluated in different test situations.

**Method**

Six simulations were conducted, where tests with one general ability and two specific abilities were considered, that is, $m = 2$. First, 82 item parameters, $\alpha$ and $\gamma$, were randomly generated from a uniform distribution so that $\alpha \sim U(0, 1)$ and $\gamma \sim U(-1, 1)$, where $\alpha$ are the item discriminations associated with the specific abilities. Then, 1,000 ability parameters, $\theta_i$, where $\theta_i = (\theta_{0i}, \theta_{1i}, \theta_{2i})'$ were simulated from $N_3(0, R_0)$. $R_0$ is a correlation matrix, which was specified to be

$$R_0 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_0 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_0 = \begin{pmatrix} 1 & 0.8 & 1 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$
\[ R_0 = \begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.6 \\ 0.6 & 0.8 & 1 \end{pmatrix}, \ R_0 = \begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.5 \\ 0.6 & 0.8 & 1 \end{pmatrix}, \text{ and } R_0 = \begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.5 \\ 0.6 & 0.8 & 1 \end{pmatrix}, \]

respectively, in the six simulations. It has to be noted that although zero correlations are unusual in practice, they were considered in the study to illustrate the extreme cases when the latent traits are not related.

Next, the response matrix \( Y \) was simulated 10 times based on the person and item parameters in each simulation. With the simulated \( Y \), Gibbs sampling was implemented for the proposed hierarchical MIRT models where 7,000 iterations were obtained with the first 2,000 as burn-in. Convergence was assessed using R statistics (Gelman, Carlin, Stern, & Rubin, 2004) as well as the Brooks-Gelman multivariate potential scale reduction factor (MPSRF; Brooks & Gelman, 1998) with multiple chains, and values close to 1 suggested that stationarity had been reached. Then, the posterior estimates were obtained as the posterior expectations of the Gibbs samples and the results for the six simulation studies are summarized in what follows.

Results and Discussion

To examine the item parameter recovery in each case, root-mean-squared differences (RMSD) between true and estimated item parameters were obtained from each replication and their averages were used to compare the two models in their parameter recoveries in the six simulations. The results are summarized in Table 1.

A close examination of the simulation results reveals that

1. Item difficulty parameters, \( \gamma \), are stable and can be recovered extremely well in different scenarios for both hierarchical MIRT models. On the other hand, item discriminations, \( \alpha \), are recovered less well.
2. Item discriminations are stable for Model 2 but not for Model 1. Specifically, they are less well recovered in Simulations 2, 5, and 6, where nonzero correlation between the specific abilities is assumed. Hence, as far as item discriminations are concerned, Model 2 works well under different simulated test situations. However, Model 1 seems to perform well only in the situation where the specific abilities are not correlated.

In general, the two hierarchical MIRT models make different assumptions on the underlying dimensional structures. They perform slightly differently in the six simulated situations. In particular, Model 2 is preferred to Model 1 as far as item parameter recovery is concerned.

Simulation Study 2

To further evaluate the performance, the proposed hierarchical MIRT models were compared with the conventional unidimensional IRT model under different
test situations using the Bayesian model choice techniques. In addition, we also compared them with the multiunidimensional IRT model to see how much they differ in the various simulated test situations.

**Method**

Five simulations were conducted, where tests with two specific abilities were considered, that is, \( m = 2 \). First, 82 item parameters, \( \alpha \) and \( \gamma \), were randomly generated from a uniform distribution so that \( \alpha \sim U(0, 1) \) and \( \gamma \sim U(-1, 1) \), where \( \alpha \) are the item discriminations associated with the specific abilities. Then, 1,000 specific ability parameters, \( \theta_i \), where \( \theta_i = (\theta_{1i}, \theta_{2i})' \) were simulated from \( N_2(\mathbf{0}, \mathbf{R}_0) \), with \( \mathbf{R}_0 \) being a correlation matrix, which was specified to be

\[
\mathbf{R}_0 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{R}_0 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, \quad \mathbf{R}_0 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}, \quad \mathbf{R}_0 = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{R}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},
\]

respectively, in the five simulations. We note that Simulation 1 is the situation where the conventional unidimensional IRT model should perform well and Simulation 5 is an extreme case where the unidimensional model does not hold.

Under each simulated scenario, the response matrix \( \mathbf{Y} \) was generated 10 times based on the item and person parameters. Then four IRT models—namely, the unidimensional, the multiunidimensional, and the two proposed hierarchical models—were implemented using the Gibbs procedure where 7,000 iterations were obtained with the first 2,000 as burn-in. Model comparisons were further carried out using

<table>
<thead>
<tr>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>.0629</td>
<td>.2069</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>.0599</td>
<td>.0510</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation 4</th>
<th>Simulation 5</th>
<th>Simulation 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>.0629</td>
<td>.1652</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>.0510</td>
<td>.0539</td>
</tr>
</tbody>
</table>

| \( \alpha \) | .0649 | .0626 | .0647 |
| \( \gamma \) | .0602 | .0539 | .0513 |
BFs and Bayesian deviances. It has to be noted that because BFs do not produce meaningful results with noninformative priors, the uniform prior distributions for item parameters specified in Section 3 were changed to be standard normal for \( \gamma_{\nu j} \) and truncated normal for \( \alpha_{\nu j} \), so that \( \gamma_{\nu j} \sim N(0,1) \) and \( \alpha_{\nu j} \sim N_{(0,\infty)}(0,1) \). The resulting posterior estimates differed from those with uniform priors for \( \xi_{\nu j} \) to two decimal places.

**Results and Discussion**

The model comparison results in the simulations were averaged over the 10 replications and are summarized in Table 2. To obtain BFs, the marginal densities \( p(y|M) \) were approximated using MCMC and are displayed in the first column of the table. Because all the likelihoods for the data were very small, the values shown in this column were some constant multiple of \( p(y|M) \), as shown in the note to the table, so that when computing BFs, the constant cancelled out. BFs are ratios of the marginal densities for comparing two models \( M_i \) and \( M_j \), that is, \( BF = (p(y|M_i))/p(y|M_j) \), which can be viewed as the odds in favor of \( M_i \) over \( M_j \). Values above one provide evidence in favor of \( M_i \), and larger values are associated with stronger evidence. Specifically, a BF beyond 100 indicates decisive evidence in favor of \( M_i \) (cf. Robert, 2001). Hence, the unidimensional model was found to be consistently worse than the multiunidimensional model or any of the two hierarchical MIRT models even when the actual latent structure was unidimensional (Simulation 1). Moreover, with relatively larger estimates of the marginal densities, there is evidence that the hierarchical MIRT Model 1 is better than the three other models in all the simulations. Taking the ratio of its marginal density with that for the unidimensional IRT model resulted in BF estimates larger than 100.

The remaining four columns of the table display the Bayesian deviance results and, in particular, the estimates averaged over the 10 replications for Bayesian DIC, the posterior expectation of the deviance (\( \bar{D} \)), the deviance of the posterior expectation (\( D(\bar{\theta}) \)), and the effective number of parameters (\( p_D \)), respectively. In all five simulations, the hierarchical MIRT Model 1 showed the smallest \( \bar{D} \) values. However, after penalizing for model complexity, that is, after taking into consideration the effective number of parameters, Bayesian DICs provided evidence for the unidimensional model in Simulation 1, where the latent dimensional structure was actually unidimensional, and the hierarchical MIRT Model 2 in Simulations 2 through 5, respectively.

With the above observations, a few remarks can be made about these simulations:

1. When the latent dimensional structure conforms to strict unidimensionality, the unidimensional IRT model is the selected model after accounting for the effective number of parameters. However, BFs support the more complicated hierarchical MIRT Model 1.
As the actual latent structure moves away from unidimensionality, the unidimensional IRT model performs less well compared with other models, even after taking into consideration the model complexity. Moreover, BFIs and Bayesian DICs do not show consistent results on the performances of the three MIRT models. Using BFIs, the first hierarchical MIRT model is shown to be better, whereas the second hierarchical MIRT model is suggested to be better using the Bayesian DIC measures. Hence, although there is no conclusive finding as to which of the hierarchical MIRT models performs better than the other, there is evidence that the two proposed MIRT models perform better than the unidimensional or the multiunidimensional IRT models when the test involves more than one common ability.

### Table 2
**Approximated Marginal Densities and Bayesian Deviances for the Four IRT Models Under Five Simulated Situations (10 Replications)**

| Simulation 1 | p(y|M) | DIC | \(\hat{D}\) | D(\(\hat{\beta}\)) | p_D |
|--------------|--------|-----|-------------|--------------------|------|
| Unidimensional model | 2.98E + 225^a | 44,676 | 43,722 | 42,768 | 954 |
| Multiunidimensional model | 4.85E + 244^a | 44,693 | 43,637 | 42,582 | 1,056 |
| Hierarchical model 1 | 4.81E + 276^a | 44,732 | 43,523 | 42,315 | 1,208 |
| Hierarchical model 2 | 2.98E + 241^a | 44,680 | 43,630 | 42,580 | 1,050 |

| Simulation 2 | p(y|M) | DIC | \(\hat{D}\) | D(\(\hat{\beta}\)) | p_D |
|--------------|--------|-----|-------------|--------------------|------|
| Unidimensional model | 3.75E + 81^a | 45,664 | 44,733 | 43,801 | 932 |
| Multiunidimensional model | 2.37E + 212^a | 45,352 | 43,995 | 42,838 | 1,357 |
| Hierarchical model 1 | 1.02E + 219^a | 45,371 | 43,959 | 42,547 | 1,412 |
| Hierarchical model 2 | 3.53E + 207^a | 45,350 | 43,997 | 42,644 | 1,353 |

| Simulation 3 | p(y|M) | DIC | \(\hat{D}\) | D(\(\hat{\beta}\)) | p_D |
|--------------|--------|-----|-------------|--------------------|------|
| Unidimensional model | 1.19E-54^a | 46,128 | 45,213 | 44,298 | 915 |
| Multiunidimensional model | 5.33E + 220^a | 45,330 | 43,832 | 42,334 | 1,498 |
| Hierarchical model 1 | 1.24E + 229^a | 45,342 | 43,815 | 42,288 | 1,527 |
| Hierarchical model 2 | 2.76E + 225^a | 45,327 | 43,832 | 42,336 | 1,496 |

| Simulation 4 | p(y|M) | DIC | \(\hat{D}\) | D(\(\hat{\beta}\)) | p_D |
|--------------|--------|-----|-------------|--------------------|------|
| Unidimensional model | 6.37E-279^a | 46,944 | 46,060 | 45,175 | 885 |
| Multiunidimensional model | 1.80E + 215^a | 45,387 | 43,795 | 42,203 | 1,592 |
| Hierarchical model 1 | 1.11E + 217^a | 45,397 | 43,788 | 42,180 | 1,609 |
| Hierarchical model 2 | 4.48E + 208^a | 45,387 | 43,794 | 42,201 | 1,593 |

| Simulation 5 | p(y|M) | DIC | \(\hat{D}\) | D(\(\hat{\beta}\)) | p_D |
|--------------|--------|-----|-------------|--------------------|------|
| Unidimensional model | 9.07E-280^b | 47,527 | 46,666 | 45,805 | 861 |
| Multiunidimensional model | 1.65E + 279^b | 45,529 | 43,908 | 42,287 | 1,621 |
| Hierarchical model 1 | 7.33E + 279^b | 45,532 | 43,907 | 42,283 | 1,624 |
| Hierarchical model 2 | 1.42E + 272^b | 45,527 | 43,908 | 42,288 | 1,619 |

Note: DIC = deviance information criteria

a. The reported values are \(p(y|M) \times \exp(22,400)\).

b. The reported values are \(p(y|M) \times \exp(22,600)\).
3. In real test situations, it is not possible to have a perfect correlation between the items in two subtests. Consequently, the advantage of the hierarchical MIRT models over the unidimensional model is more obvious for real test data.
4. It has to be noted that the multidimensional IRT model performs relatively similar to the proposed two hierarchical IRT models. Both BF s and Bayesian DIC s indicate that it is the second best model when the test is not strictly unidimensional.

A Real Data Example

As an illustration, the proposed models were subsequently implemented on a subset of CBASE English subject data. In real test situations such as this, the actual latent structure is not readily known. Hence, model comparison is used to determine which candidate model provides a relatively better representation of the data.

Method

The overall CBASE exam contains 41 English multiple choice items, with the first 16 items in writing cluster and the remaining 25 in the reading/literature cluster. The data used in this study were from college students who took the LP form of CBASE in years 2001 and 2002. After removing all those who attempted the exam multiple times and removing missing responses, a sample of 1,231 examinees was randomly selected. To assess the goodness of fit, the proposed hierarchical MIRT models were compared with the conventional unidimensional IRT model as well as the multidimensional model. The four candidate models were implemented on the CBASE English data using the Gibbs sampling procedure, where 7,000 iterations were obtained with the first 2,000 set as burn-in. R statistics (Gelman et al., 2004) and the Brooks-Gelman MPSRF (Brooks & Gelman, 1998) were used to assess convergence, and they were found to be close to 1, suggesting that stationarity had been reached within the simulated Monte Carlo chains for the models. Then, model comparisons were carried out using BF s and Bayesian DIC s.

Results and Discussion

Table 3 displays the results for BF s and Bayesian deviances. The first column lists the approximated marginal densities \( p(y|M) \) for the four candidate models. As a BF beyond 100 indicates decisive evidence in favor of the model on the numerator, the two hierarchical MIRT models were found to be better than the unidimensional or the multidimensional IRT models. In addition, among the two hierarchical MIRT models, Model 1 was shown to be better for the sample data. On the other hand, there was much evidence against the unidimensional IRT model
when comparing it with either the multiunidimensional model or the two hierarchical MIRT models.

Table 3 also displays the Bayesian deviance results, where smaller values indicate better model fit. Among the four candidate IRT models, the proposed hierarchical MIRT Model 2 and the multiunidimensional model had smaller DIC values. Moreover, the hierarchical MIRT Model 1 had the largest effective number of parameters ($p_D = 1,516$) and was more complicated than the other three models. Hence, after penalizing for model complexity, it was shown to be only better than the unidimensional IRT model. Furthermore, taking into consideration model complexity, Bayesian DICs did not show evidence for the unidimensional IRT model either.

Therefore, with Bayesian model checking techniques, the four candidate IRT models were evaluated as to which model provided the best representation of the $CBASE$ data. The results from BF$s$ and Bayesian deviances showed evidence in favor of the proposed two hierarchical models, which fit the data better than the conventional unidimensional model. However, as BF$s$ and Bayesian DICs suggested slightly different results, it is not clear which hierarchical MIRT model is the “best”. The unidimensional model provided a relatively worse description of the data. Hence, for the $CBASE$ English data, the model comparison results did not support the more stringent unidimensionality assumption.

### Discussion and Conclusion

In conclusion, this study incorporates the latent structure of second-order factor models into the IRT framework and proposes two hierarchical MIRT models that focus on the person–item interaction. These models, though making different assumptions on the relationship between the general ability and specific abilities, are found to describe test data similarly well in different test situations, with one being slightly better. Simulation studies indicate that (a) in parameter recovery,
the hierarchical Model 2 performs well in all test situations, whereas the hierarchical MIRT Model 1 recovers item parameters less well when the specific ability dimensions are correlated, and (b) in model comparisons, especially when the test is not actually unidimensional, the two hierarchical models are shown to be better than the conventional unidimensional model and, in some cases, the multi-
unidimensional model, using different Bayesian model selection criteria.

As far as the CBASE data is concerned, the proposed hierarchical MIRT models provide relatively better descriptions of the data than the conventional unidimensional model or even the multiunidimensional model. Moreover, the unidimensional model describes the CBASE data most poorly compared with models with multiple dimensions. Therefore, using the unidimensional model for the CBASE English test is not supported. The actual dimensionality for the test proves to be more complex than a single latent trait. The model with one general and two specific ability dimensions is closer to reality among the models considered. All test items measure two specific abilities, which are further related to an underlying general ability dimension. The latter estimates the examinee’s overall English ability, which has the same interpretation as the ability dimension assumed in the unidimensional model. Indeed, their estimates are close for the CBASE English data, as can be seen in Figure 1 (left panels). On the other hand, the unidimensional model consistently underestimates the posterior standard deviation (right panels of Figure 1), simply because the single latent trait fails to capture the extra variability in the specific ability dimensions. Moreover, a comparison between the two scatter plots on the right panels of the figure reveals noticeable dissimilarity in the variability of the general ability in the two hierarchical models. This actually reflects the fundamental difference between the two MIRT models, which is comparable to the difference between factor analysis and principal components. In effect, the general ability in Model 1 is a latent trait accounting for common variance in the specific traits. Its score provides an estimate of the underlying latent general trait. On the other hand, the general ability in Model 2 is comparable to a principal component accounting for a maximal amount of variance of the specific traits. Its score can be interpreted as the weighted average score based on the estimated specific latent traits.

It has to be noted that with the same probability function, the proposed hierarchical MIRT models and the multiunidimensional IRT model differ only in the inclusion of the general ability dimension. Figure 2 illustrates their differences graphically. The multiunidimensional model (Figure 2a) assumes that specific ability dimensions, correlated or uncorrelated, underlie all test items. The proposed hierarchical MIRT models, on the other hand, assume that each specific ability is either a linear function of the general ability (Figure 2b) or linearly combines to form the general ability (Figure 2c), so that test items measure the general ability dimension indirectly through the specific ability dimensions. Therefore, these models make different assumptions,
and they should be adopted based on different beliefs concerning the underlying dimensional structure of a test.

Moreover, the fact that the general ability underlies the latent specific abilities makes the proposed hierarchical MIRT models applicable only in the test situation where the general ability is closely related to each specific ability, for little information can be utilized to make an inference on the general ability if otherwise. Further studies are needed to develop such models in other test situations.

In the current study, BFs were approximated because of the difficulty with the exact analytic evaluation for complicated hierarchical Bayesian models. The harmonic mean of the likelihood, which is used to approximate the marginal likelihood of the data using MCMC methods, converges to the correct value as the chain
length goes to infinity. However, it does not satisfy a Gaussian central limit theorem, because the model parameter may take an occasional value with small likelihood, which results in a large effect on the final result. Future studies should adopt more accurate methods that are based on the estimation of marginal likelihoods, such as the Chib’s method (Chib, 1995; Chib & Jeliazkov, 2001) or the bridge sampling method (Meng & Schilling, 2002; Meng & Wong, 1996).

References


