Behaviormetrika
Vol.36, No.1, 2009, 27–48

BAYESIAN IRT MODELS INCORPORATING GENERAL AND SPECIFIC ABILITIES

Yanyan Sheng* and Christopher K. Wikle**

IRT-based models with a general ability and several specific ability dimensions are useful. Studies have looked at item response models where the general and specific ability dimensions form a hierarchical structure. It is also possible that the general and specific abilities directly underlie all test items. A multidimensional IRT model with such an additive structure is proposed under the Bayesian framework. Simulation studies were conducted to evaluate parameter recovery as well as model comparisons. A real data example is also provided. The results suggest that the proposed additive model offers a better way to represent the test situations not realized in existing models.

1. Introduction

Unidimensional item response theory (IRT) models are useful when tests are designed to measure only one ability that may be explained by one latent trait or a specific combination of traits. However, psychological processes have consistently been found to be more complex and an increasing number of educational measurements assess an examinee on more than one latent trait. With regard to this, allowing separate inferences to be made about an examinee for each distinct ability dimension being measured, multidimensional IRT (MIRT) models have shown promise for dealing with such complexity in educational and psychological measurement (Ackerman, 1993; Reckase, 1997). With the use of Bayesian estimation procedures, different multidimensional models involving continuous latent traits have been developed, including MIRT models where each item measures multiple abilities (Béguin & Glas, 2001), multi-unidimensional IRT models where multiple specific ability dimensions are involved in one test with each item measuring only one of them (e.g., Lee, 1995; Sheng & Wikle, 2007), and hierarchical MIRT models where each item measures a specific ability, which is further related to an underlying general ability (Sheng & Wikle, 2008).

The hierarchical MIRT models proposed by Sheng and Wikle (2008) have been shown to perform better than the traditional unidimensional IRT model. However, they assume that the general and specific ability dimensions form a hierarchical structure so that each specific ability is either a linear function of the general ability or linearly combines to form the general ability. This structure requires the actual general ability to be correlated with the specific abilities. Otherwise, little information can be drawn to make inference on the underlying general ability. Hence, hierarchical models are not applicable in all educational and psychological test situations. In this paper, we propose another IRT-based model in-

Key Words and Phrases: item response theory, additive MIRT model, hierarchical MIRT model, unidimensional model, multi-unidimensional model, MCMC, Bayesian model choice.

* Department of Educational Psychology & Special Education, Southern Illinois University-Carbondale, Wham 223 - Mail Code 4618, Carbondale, IL 62901, USA. E-mail: ysheng@siu.edu
** Department of Statistics, University of Missouri

corporating both general and specific ability dimensions under the Bayesian framework so that the general ability and the specific ability dimensions form an additive structure, i.e., each item measures a general and a specific ability directly. We call this the additive MIRT model and believe it is not restricted to situations where the general and specific ability dimensions are correlated. It has to be noted that when referring to the specific ability, we do not limit ourselves to what Spearman posited in his two-factor theory (Spearman, 1904), where specific abilities, or more precisely, specific factors can be thought of "nuisance" factors (Segall, 2001, p.80) that are not correlated among themselves or with the general factor. Rather, given the number of mental ability theories that have emerged since the early twentieth century, and the difficulty in coming up with an unanimously accepted definition or classification of the non-general abilities, we consider here also cases where specific ability is the cognitive process needed for an individual subtest that may be related to the overall trait (e.g., reading comprehension ability vs. ability for reading, writing, and listening), or may be related to those required for other subtests (such as reading comprehension ability vs. writing ability). Hence, the additive MIRT model is compared with hierarchical MIRT models under various situations where the underlying abilities have different levels of correlation. Further, to illustrate the Gibbs sampling procedure for the proposed model, a subset of *College Basic Academic Subjects Examination* (*CBASE*; Osterlind, 1997) *English* subject data is examined.

The remainder of the paper is organized as follows. Section 2 reviews the conventional unidimensional and multi-unidimensional models as well as the hierarchical MIRT models from Sheng and Wikle (2008), while Section 3 describes the proposed additive MIRT model in the Bayesian framework. The Gibbs sampling procedure is also illustrated in this section. Section 4 illustrates the Bayesian model selection techniques. To evaluate model performance, simulation studies were conducted to recover parameters as well as to compare the proposed additive model with the hierarchical models under different test situations using Bayesian model selection procedures. The results from the simulation studies are summarized in Sections 5 and 6. Section 7 gives an example where the proposed model is implemented on a subset of *CBASE English* subject data and Bayesian model selection procedures are subsequently performed to compare this model with the conventional IRT models as well as the hierarchical models. Finally, a few summary remarks are given in Section 8.

2. IRT models

In this paper, we focus primarily on two-parameter normal ogive (probit) models.

2.1 Unidimensional IRT model

The unidimensional IRT model provides the simplest framework for modeling the person-item interaction by assuming one ability dimension. Suppose a test consists of k binary-response items (e.g., multiple-choice items), each measuring a single unified ability, θ. Define y_{ij} as

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

Then, the probability of person i obtaining the 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 α_j is a scalar parameter describing the item discrimination, γ_j is associated with item difficulty β_j such that $\gamma_j = \alpha_j \ \beta_j$, and θ_i is a scalar ability parameter.

2.2 Multi-unidimensional IRT model

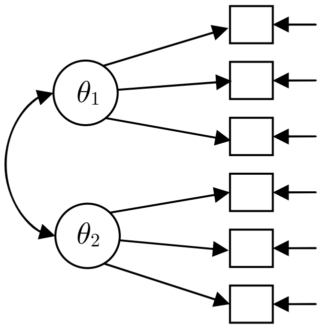
Multi-unidimensional models allow separate inferences to be made about an examinee for each distinct dimension being measured by a subtest question (Sheng & Wikle, 2007). Consider a K-item test consisting of m subtests, each containing k_v binary-response items that measure one ability dimension. With 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:

$$P(y_{vij} = 1|\theta_{vi}, \alpha_{vj}, \gamma_{vj}) = \Phi(\alpha_{vj} \theta_{vi} - \gamma_{vj}) = \int_{-\infty}^{\alpha_{vj} \theta_{vi} - \gamma_{vj}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt, \quad (2)$$

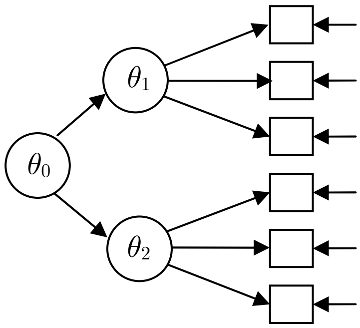
where α_{vj} and θ_{vi} are scalar parameters representing the item discrimination and the examinee ability in the v-th ability dimension, and γ_{vj} is a scalar parameter indicating the location in that dimension where the item provides maximum information.

2.3 Hierarchical MIRT models

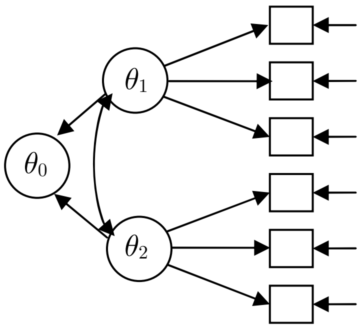
Incorporating the latent structure of second-order factor models (Schmid & Leiman, 1957) into IRT framework, the hierarchical MIRT model (Sheng & Wikle, 2008) assumes the same probability function as that of the multi-unidimensional models specified in (2). They specify a hierarchical structure so that each specific ability either 1) is a linear function of the general ability (Figure 1b) so that $\theta_{vi} \sim N(\delta_v \theta_{0i}, 1)$, where θ_{0i} is the i-th examinee ability parameter corresponding to the overall test, or 2) linearly combines to form the general ability (Figure 1c) so that $\theta_{0i} \sim N(\sum_v \lambda_v \theta_{vi}, 1)$. In this paper, we refer to these two formulations as hierarchical MIRT model 1 and hierarchical MIRT model 2, respectively. As Figure 1 shows, they can be considered as extensions of the multi-unidimensional model (Figure 1a), with more complicated underlying dimensional structures.



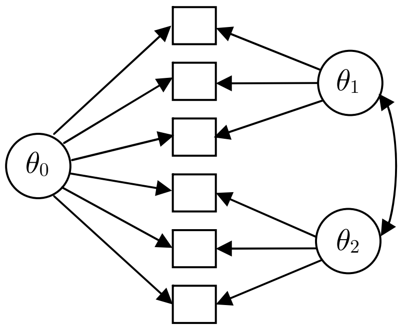
(a) multi-unidimensional model



(b) hierarchical MIRT model 1



(c) hierarchical MIRT model 2



(d) additive MIRT model

Figure 1: Graphical illustrations of the multi-unidimensional IRT model, the two hierarchical MIRT models and the proposed additive MIRT model. Circles represent latent traits, and squares represent observed items.

3. The proposed Bayesian IRT model

3.1 Additive MIRT model

The proposed additive MIRT model differs from the hierarchical MIRT models in that the general ability directly affects the examinee's response to a test item (Figure 1d). In other words, the latent trait dimensions form an additive structure.

For a K-item test containing m subtests, each with k_v binary-response items, where $v = 1, \ldots, m$, y_{vij} is the response for the i-th examinee on the j-th item of the v-th subtest. With a two-parameter probit model, we define the probability function $p_{vij} = P(y_{vij} = 1)$ as

$$P(y_{vij} = 1) = \Phi(\alpha_{0vj}\theta_{0i} + \alpha_{vj}\theta_{vi} - \gamma_{vj}), \tag{3}$$

where θ_{vi}, θ_{0i}, and γ_{vj} are as defined in the previous section, α_{0vj} is the j-th item discrimination parameter associated with the general ability, θ_{0i}, and α_{vj} is the item discrimination associated with the specific ability, θ_{vi}. Hence, the probability of answering an item correctly is assumed to be determined directly by two latent traits—a general and a specific one.

One should note the similarity of this formulation with that of Bradlow, Wainer, and

Wang's (1999) so-called "testlet" model, whose systematic component takes the form $\alpha_j\theta_{0i} - \gamma_j - \alpha_j\theta_{i(v)}$, where $\theta_{i(v)} \sim N(0, \sigma^2)$. It can be shown that the testlet model is a special case of the proposed additive MIRT model where $\alpha_{0vj} = \alpha_{vj}$. That is, if expressed in our context, each item differentiates between examinees in their general and specific abilities equally, although in the opposite directions. Moreover, the proposed model allows one to specify a different distribution for α_0, α, or γ for each subtest, whereas the testlet model does not. The latter is hence limited in the situations when, for instance, it is believed that items in a particular subtest are supposed to have very different characteristics than those in other subtests. Finally, the testlet model assumes zero correlations among the specific abilities, whereas the additive model, as is illustrated in the following section, models their interdependence by introducing a covariance structure for their mean vectors $\boldsymbol{\mu}_i$. This further illustrates that the additive model is more general and thus offers more flexibility than the testlet model. Indeed, the result of the simulation study shown in Appendix A provides empirical evidence that the testlet model does not work as well as the additive model when α_0 and α are not constrained to be the same, and hence is limited in situations where its model assumptions are violated. Given this, the testlet model was not considered in the analyses presented here.

Additionally, one may reformulate the hierarchical MIRT model 1 so that its systematic component takes the form $\alpha_{vj}\delta_v\theta_{0i} + \alpha_{vj}\varepsilon_{vi} - \gamma_{vj}$, where $\varepsilon_{vi} \sim N(0, 1)$, and claim that it is a constrained version of the additive model. However, the two models differ fundamentally in that their parameters, θ_{vi} and ε_{vi}, have different interpretations. Specifically, θ_{vi} in the additive model denotes the specific ability for the v-th subtest, which can be correlated with other specific abilities, or with the general ability, θ_{0i}, as is illustrated in the following section. Nevertheless, ε_{vi} in the hierarchical model 1 denotes independent random error specific for the v-th subtest, and has a zero correlation with the general ability, θ_{0i}.

We denote each examinee's abilities for all items as $\boldsymbol{\theta}_i = (\theta_{0i}, \theta_{1i}, \theta_{2i}, \ldots, \theta_{mi})'$, vectors of $m+1$ ability parameters and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)'$. Also, denote $\boldsymbol{\xi}_{vj} = (\alpha_{0vj}, \alpha_{vj}, \gamma_{vj})'$ the vector of item parameters for the j-th item of the v-th subtest and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)'$, where $\boldsymbol{\xi}_v = (\boldsymbol{\xi}_{v1}, \ldots, \boldsymbol{\xi}_{vk_v})'$. With the assumption of local independence, i.e., conditional on $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ the responses are independent, the joint probability function of $\mathbf{y}$, where $\mathbf{y} = [y_{vij}]_{n\times K}$ is

$$P(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{v=1}^{m}\prod_{i=1}^{n}\prod_{j=1}^{k_v} p_{vij}^{y_{vij}}(1-p_{vij})^{1-y_{vij}}, \tag{4}$$

where p_{vij} is as specified in (3).

3.2 Model specification

Assume that the prior distribution of $\boldsymbol{\theta}_i$, $i = 1, \ldots, n$, is multivariate normal (MVN) with mean $\boldsymbol{\mu}_i$, where $\boldsymbol{\mu}_i = (\mu_{0i}, \mu_{1i}, \ldots, \mu_{mi})'$, and covariance matrix $\mathbf{I}$, the identity matrix, so the prior probability density function for the abilities is

$$\varphi_{m+1}(\boldsymbol{\theta}_i; \boldsymbol{\mu}_i, \mathbf{I}) = (2\pi)^{-\frac{m+1}{2}} \exp\{-(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)'(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)/2\}. \tag{5}$$

Note that any unconstrained covariance matrix can be adopted for the prior distribution. However, the identity matrix is adopted here to set a strong prior for the latent traits so as to get around the model indeterminacy problem (see Lee, 1995 for a statement of the problem). Also, the hyperparameters $\boldsymbol{\mu}_i$,i=1, ..., n, are assumed to be independent MVN with mean $\mathbf{0}$, where $\mathbf{0} = (0, \ldots, 0)'$, and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is assumed to have an inverse-Wishart distribution $\boldsymbol{\Sigma} \sim W^{-1}(\mathbf{I}, m{+}1)$. So the density function for $\boldsymbol{\mu}_i$ is

$$\varphi_{m+1}(\boldsymbol{\mu}_i; \mathbf{0}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{m+1}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i/2\}. \tag{6}$$

It should be noted again that the correlations between $\theta_{0i}, \theta_{1i}, \theta_{2i}, \ldots$, and θ_{mi} are modeled through the common mean structure so that the dependence in hyperparameters $\boldsymbol{\mu}_i$ with the use of an unconstrained covariance matrix $\boldsymbol{\Sigma}$ leads to dependence in the ability parameters $\boldsymbol{\theta}_i$. We set conjugate normal priors for $\boldsymbol{\xi}_{vj}$, $v = 1, \ldots, m$, $j = 1, \ldots, k_v$ so that $\alpha_{0vj} \sim N_{(0,\infty)}(0, 1)$, $\alpha_{vj} \sim N_{(0,\infty)}(0, 1)$ and $\gamma_{vj} \sim N(0, 1)$, and assume the prior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are independent.

Hence, by introducing an augmented continuous variable $\mathbf{Z}$ (Albert, 1992; Tanner & Wong, 1987) such that $Z_{vij} \sim N(\eta_{vij}, 1)$, where $\eta_{vij} = \alpha_{0vj}\theta_{0vi} + \alpha_{vj}\theta_{vi} - \gamma_{vj}$ and $y_{vij} = \begin{cases} 1, & if \quad Z_{vij} > 0 \\ 0, & if \quad Z_{vij} \leq 0 \end{cases}$, the joint posterior distribution of ($\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\mathbf{Z}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) is then

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Z}, \boldsymbol{\Sigma}, \boldsymbol{\mu}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{Z})p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\xi})p(\boldsymbol{\theta}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}). \tag{7}$$

The full conditional distributions can be derived in closed form, as shown in Appendix B. Hence, the Gibbs sampler can be adopted to iteratively update samples $\mathbf{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from their respective full conditionals in (9), (11), (13), (15) and (17), with starting values $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$. The collection of all these simulated draws from $p(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ are then used to summarize the posterior density of item parameters $\boldsymbol{\xi}$ and ability parameters $\boldsymbol{\theta}$ and can be used to compute quantiles, moments and other summary statistics. As with standard Monte Carlo, with large enough samples, the posterior means of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are considered as estimates of the true parameters. It has to be noted that this model specification, however, does not directly model the correlation between the latent abilities. In the situations where the inter-trait correlations are of interest, one has to estimate them indirectly via correlating the posterior estimates of the ability parameters.

4. Bayesian model choice techniques

From the frequentist's perspective, it is natural to compare several models using likelihood ratio tests or other information criteria. Likewise, in the Bayesian framework, model comparison/selection is made possible with several criteria, among which, Bayes factors, Bayesian deviance and posterior predictive model checks are to be considered in this study.

4.1 Bayes factor

When a set of s different Bayesian hierarchical models $M_1, \ldots, M_s$ are considered, the

Bayes factor for comparing two models M_i and M_j is defined as $BF = \dfrac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)}$, where $p(\mathbf{y}|M) = \int_{\vartheta} L(\mathbf{y}|\vartheta)p(\vartheta|M)d\vartheta$ is the marginal probability of the data $\mathbf{y}$ (also referred to as the prior mean of the likelihood) with ϑ denoting all model parameters, and $p(\vartheta|M)$ is the prior density for the unknown parameters under the specific model M. This is the Bayesian analogue of the likelihood ratio between two models, and describes the evidence provided by the data in favor of M_i over M_j. The Bayes factors allow comparison of non-nested models and ensure consistent results for model comparisons, but are usually difficult to calculate due to 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, etc. 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 can be unstable, research shows that it often produces results that are accurate enough for interpreting the Bayes factors (e.g., Carlin & Chib, 1993) and therefore it was used in this study.

To estimate the marginal density, one can draw MCMC samples of the parameters, $\vartheta^{(1)}, \dots, \vartheta^{(G)}$, so that $p(\mathbf{y}|M)$ is approximated as $\left\{ \dfrac{1}{G} \sum_{g=1}^{G} L(\mathbf{y}|\vartheta^{(g)})^{-1} \right\}^{-1}$. This is defined as the harmonic mean of the likelihood values (Newton & Raftery, 1994). In addition, Aitkin (1991) proposed a posterior Bayes factor $PBF = \dfrac{p^*(\mathbf{y}|M_i)}{p^*(\mathbf{y}|M_j)}$ for Bayesian models with improper priors, where $p^*(\mathbf{y}|M) = \int_{\vartheta|\mathbf{y}} L(\mathbf{y}|\vartheta)p(\vartheta|\mathbf{y}, M)d\vartheta$ is the posterior mean of the likelihood. To approximate this marginal density, one again uses the posterior samples so that $p^*(\mathbf{y}|M) = \dfrac{1}{G}\sum_{g=1}^{G} L(\mathbf{y}|\vartheta^{(g)})$. In this study, we considered both Bayes factor (BF) and posterior Bayes factor (PBF) although all model priors were chosen to be proper.

4.2 Bayesian Deviance

The Bayesian deviance information criterion (DIC) was introduced by Spiegelhalter, Best, Carlin, and van der Linde (2002) who generalized the classical information criteria to one that is based on the posterior distribution of the deviance. This criterion is defined as $DIC = \bar{D} + p_D$, where $\bar{D} \equiv E_{\vartheta|\mathbf{y}}(D) = E(-2\log L(\mathbf{y}|\vartheta))$ is the posterior expectation of the deviance (with L being the likelihood function), and $p_D = E_{\vartheta|\mathbf{y}}(D) - D(E_{\vartheta|\mathbf{y}}(\vartheta)) = \bar{D} - D(\bar{\vartheta})$ is the effective number of parameters (Carlin & Louis, 2000). In addition, let $D(\bar{\vartheta}) = -2\log(L(\mathbf{y}|\bar{\vartheta}))$, where $\bar{\vartheta}$ is the posterior mean. To compute Bayesian DIC, MCMC samples of the parameters, $\vartheta^{(1)}, \dots, \vartheta^{(G)}$, can be drawn from the Gibbs sampler, then $\bar{D}$ his approximated as $\bar{D} = \dfrac{1}{G}(-2\log\prod_{g=1}^{G} L(\mathbf{y}|\vartheta^{(g)}))$. Generally more complicated models tend to provide better fit. Hence, penalizing for number of parameters makes DIC a more reasonable measure to use. However, unlike the Bayes factor, DIC is not invariant to parameterization and sometimes can produce unrealistic results.

4.3 Posterior predictive model checks

Among the methods proposed for model checking, posterior predictive checking is easy to carry out and interpret in spite of its limitation in being conservative (Sinharay & Stern, 2003). The basic idea is to draw simulated values from the posterior predictive distribution of replicated data, $\mathbf{y}^{\mathrm{rep}}$, $p(\mathbf{y}^{\mathrm{rep}}|\mathbf{y}) = \int p(\mathbf{y}^{\mathrm{rep}}|\vartheta)p(\vartheta|\mathbf{y})d\vartheta$, and compare them to the observed data $\mathbf{y}$. If the model fits, then replicated data generated under the model should look similar to the observed data. A test statistic $\mathrm{T}(\mathbf{y}, \vartheta)$ has to be chosen to define the discrepancy between the model and the data. If there are L simulations from the posterior distribution of ϑ, one $\mathbf{y}^{\mathrm{rep}}$ can be drawn from the predictive distribution for each simulated ϑ so there are L draws from the joint posterior distribution $p(\mathbf{y}^{\mathrm{rep}}, \vartheta|\mathbf{y})$. It is then easy to compare the realized test statistics $\mathrm{T}(\mathbf{y},\ \vartheta^l)$ with the predictive test statistics $\mathrm{T}(\mathbf{y}^{\mathrm{rep}},\ \vartheta^l)$ by plotting the pairs on a scatter plot. Alternatively, one can calculate the probability or posterior predictive p-value (PPP-value) (Sinharay, Johnson, & Stern, 2006) that the replicated data could be more extreme than the observed data: $p_{\mathrm{B}} = \Pr(T(\mathbf{y}^{\mathrm{rep}}, \vartheta^l) \geq T(\mathbf{y}, \vartheta^l)|\mathbf{y})$.

5. Parameter recovery

In the proposed model, each test item is assumed to measure two ability dimensions, namely, a general and a specific ability dimension, directly. This is reflected in the probability function of the model defined in (3). The additive nature of the latent traits in the model leads to a potential problem of indeterminancy when item and person parameters are estimated simultaneously. In the Bayesian framework, although some strong informative priors are specified for the ability parameters to help the convergence of Markov chains, it is still uncertain how the Bayesian additive MIRT model performs in various scenarios. Hence, a series of simulation experiments was carried out to evaluate the model in item parameter recovery.

5.1 Methodology

Five simulations were conducted, where tests with one general ability and two specific abilities were considered, i.e., $m = 2$. For each simulation, a 1,000-by-41 dichotomous response data matrix $\mathbf{y}$ was simulated 10 times from the additive model defined in (3). To generate $\mathbf{y}$, α_0, α, and γ were randomly drawn from uniform distributions so that $\alpha_0 \sim U(0,1)$, $\alpha \sim U(0,1)$, $\gamma \sim U(-1,1)$, and $\boldsymbol{\theta}_i$ were simulated from $N_3(\mathbf{0}, \mathbf{R}_0)$, where $\mathbf{R}_0$ is a correlation matrix and was specified to be $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0.8 & 1 & \\ 0.6 & 0 & 1 \end{pmatrix}$,

Table 1: Average RMSD between the actual and estimated item parameters for Gibbs sampling with the two additive models under five simulated scenarios (10 replications).

	Simulation 1	Simulation 2	Simulation 3	Simulation 4	Simulation 5
Known prior					
$\hat{\alpha}_0$	0.0808	0.1089	0.1336	0.1204	0.0894
$\hat{\alpha}$	0.0797	0.1032	0.1479	0.1387	0.0966
$\hat{\gamma}$	0.0606	0.0571	0.0642	0.0675	0.0538
Proposed					
$\hat{\alpha}_0$	0.0794	0.3779	0.0996	0.3093	0.3065
$\hat{\alpha}$	0.0772	0.1813	0.2019	0.1722	0.1738
$\hat{\gamma}$	0.0608	0.06	0.0637	0.0699	0.0542

$\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0.6 & 1 \end{pmatrix}$, $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0.8 & 1 & \\ 0 & 0 & 1 \end{pmatrix}$, and $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0.5 & 1 & \\ 0.5 & 0.5 & 1 \end{pmatrix}$ in the five simulations, respectively. 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.

For each simulated $\mathbf{y}$, the Gibbs sampler was implemented to fit two Bayesian additive models. They differed only in the specification of the prior distribution for $\boldsymbol{\theta}_i$ so that one model assumed $\boldsymbol{\theta}_i \sim N_3(\mathbf{0}, \mathbf{R}_0)$, where $\mathbf{R}_0$ is the actual correlation matrix used to generate $\boldsymbol{\theta}_i$ in each simulation, and the other assumed $\boldsymbol{\theta}_i \sim N_3(\boldsymbol{\mu}_i, \mathbf{I})$, where $\boldsymbol{\mu}_i \sim N_3(\mathbf{0}, \boldsymbol{\Sigma})$. It has to be noted that the former, referred to as the model with known prior, would help detect any computational problem in the implementation of the Gibbs sampler, and the latter is exactly the proposed model. Each implementation was carried out with a run length of 7,000 iterations and a burn-in period of 2,000. Convergence was assessed using the Gelman-Rubin R statistic (Gelman, Carlin, Stern, & Rubin, 2004) with multiple chains and values close to 1 suggesting that stationarity had been reached. Hence, the posterior estimates were obtained as the posterior expectations of the Gibbs samples and the results for the five simulations are summarized as follows.

5.2 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 with respect to parameter recoveries in the five simulations. The results are summarized in Table 1.

A close examination of the results in Table 1 reveals that:

1) As expected, the model with the known prior performs relatively better in all the five simulations. This further confirms that no computational problem occurred during the implementation of the Gibbs sampling procedure.
2) With the proposed model, the location parameters γ are always well recovered and hence they are not affected by various actual structures existing in the latent traits.

On the contrary, α_0 and α are affected, and this can be explained by the fact that they are slopes for the corresponding abilities in the model. It is further noticed that when there is no correlation between the general ability and each specific ability, α_0 and α are recovered well, as shown in simulations 1 and 3. However, when the general ability is correlated with any of the specific abilities, the slopes, especially α_0, are less well recovered. Furthermore, a comparison between simulations 2, 4, and 5 indicates that the higher the correlations between θ_0 and θ_1 and/or θ_2, the less well the item parameters are recovered.

In general, the additive model implemented with Gibbs sampling is found to work well when there is no or low correlation between the general ability and each specific ability. This is because the model specifies a generalized linear function of the general ability and a specific ability. The collinearity problem, i.e., high correlations between the general ability and specific abilities, affects the accuracy of parameter estimation.

6. Model comparison

To further evaluate the performance, the proposed additive model was compared with the hierarchical MIRT models under various simulated test situations using the Bayesian model choice techniques.

6.1 Methodology

To compare the two types of MIRT models, eight simulations were conducted, where tests with one general ability and two specific abilities were considered, i.e., $m = 2$. Four of the eight simulations assumed that the hierarchical MIRT model was true, and the other four assumed that the additive MIRT model was true. Dichotomous item responses of 1,000 persons to 41 items were simulated so that, in the four simulations where the hierarchical model was true, the responses y_{vij} were generated from the probabilities as defined in (2), where $\alpha_{vj} \sim U(0, 1)$, $\gamma_{vj} \sim U(-1, 1)$. On the other hand, in the four simulations where the additive model was true, y_{vij} were simulated from the probabilities as defined in (3), where $\alpha_{0vj} \sim U(0, 1)$, $\alpha_{vj} \sim U(0, 1)$ and $\gamma_{vj} \sim U(-1, 1)$. Under both of the two conditions described previously, the ability parameters $\boldsymbol{\theta}_i$ were simulated from $N_3(\mathbf{0}, \mathbf{R}_0)$, where $\mathbf{R}_0$ is a correlation matrix and was specified to be $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0.8 & 1 & \\ 0.6 & 0 & 1 \end{pmatrix}$, $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0.6 & 1 \end{pmatrix}$, and $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0.8 & 1 & \\ 0.6 & 0.5 & 1 \end{pmatrix}$ in the four simulations, respectively.

With the simulated responses, the hierarchical and additive MIRT models were implemented using Gibbs sampling where 7,000 iterations were obtained with the first 2,000 as burn-in, which was sufficient for the chains to reach stationarity. Ten replications were used and the posterior expectations of the Gibbs samples were used to obtain the posterior estimates necessary to derive Bayes factors as well as Bayesian deviance results.

6.2 Results and Discussion

The model comparison results in each simulation were averaged over the ten replications and are summarized in Table 2 and Table 3. To obtain Bayes factors, the marginal densities $p(\mathbf{y}|M)$ and $p^*(\mathbf{y}|M)$ were approximated using MCMC and are displayed in the first two columns of the tables. Since all the likelihoods for the simulated data were very small, the values shown in the two columns are a constant multiple of $p(\mathbf{y}|M)$ or $p^*(\mathbf{y}|M)$, as is noted below the tables. However, note that when computing Bayes factors, this constant cancelled out. Bayes factors and posterior Bayes factors are ratios of the marginal densities for comparing two models M_i and M_j, i.e., $BF = \frac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)}$, $PBF = \frac{p^*(\mathbf{y}|M_i)}{p^*(\mathbf{y}|M_j)}$, and values larger than 1 provide evidence in favor of M_i to M_j. As a BF or PBF greater than 100 indicates decisive evidence in favor of M_i (cf., Robert, 2001), the additive MIRT model was found to be consistently better than the two hierarchical MIRT models even when the actual latent dimension conformed to the hierarchical structure. Taking the ratio of its marginal density with that for any of the other two models resulted in BF or PBF estimates greater than 100.

The remaining four columns of the tables show the Bayesian deviance results, and in particular, the estimates averaged over the ten replications for the Bayesian DIC, the posterior expectation of the deviance ($\bar{D}$), the deviance of the posterior expectation ($D(\bar{\vartheta})$) values, and the effective number of parameters (p_D), respectively. The proposed additive MIRT model shows consistently smaller DIC, $\bar{D}$, and $D(\bar{\vartheta})$, compared with the two hierarchical MIRT models. Since small deviance values indicate better model fit, the additive MIRT model is shown to provide a better description of the simulated data in all the simulations, even after penalizing for model complexities, i.e., the effective number of parameters. Hence, the Bayesian deviance results were consistent with the results using Bayes factors in model comparisons.

After a close examination and comparison of the values shown in the two tables, a few remarks can be drawn from these results:

1. No matter what the actual condition is, either when the additive MIRT model is true or when the hierarchical MIRT model is true, the additive MIRT model always outperforms the hierarchical MIRT models and thus provides a better description of the simulated data. The degree of this improved performance is much higher when the latent ability dimensions form an additive structure. The fact that the additive model works better even when the hierarchical model is true poses a situation worth noting. This may be due to the reason that each item is related to the general ability directly in the additive model whereas they are related indirectly in the hierarchical model. However, further analysis has to be conducted to investigate this result.
2. The effective number of parameters (p_D) displayed in the last column of the two tables gives rise to an interesting finding. When the hierarchical model is true, p_D

Table 2: Approximated marginal densities and Bayesian deviance estimates (averaged over 10 replications) for the three MIRT models under 4 simulated situations when the hierarchical model is true.

		$p(\mathbf{y}\|M)^1$	$p^*(\mathbf{y}\|M)^2$	DIC	$\bar{\text{D}}$	$D(\bar{J})$	p_D
Simualtion 1							
	Additive model	9.36E+126	3.06E+104	45212	43601	41989	1611
	Hierarchical model 1	3.06E+56	5.58E+13	45542	43925	42307	1618
	Hierarchical model 2	2.94E+42	3.51E+10	45536	43925	42313	1611
Simualtion 2							
	Additive model	5.01E+140	7.87E+132	45193	43602	42011	1591
	Hierarchical model 1	5.22E+80	2.86E+42	45538	43921	42304	1617
	Hierarchical model 2	1.18E+79	5.39E+46	45532	43921	42309	1611
Simualtion 3							
	Additive model	4.91E+91	4.19E+89	45192	43711	42230	1481
	Hierarchical model 1	6.82E+45	27E+09	45545	44021	42497	1524
	Hierarchical model 2	2.39E+38	2.14E+07	45525	44038	42552	1487
Simualtion 4							
	Additive model	6.59E+81	79E+97	45199	43646	42093	1553
	Hierarchical model 1	6.00E+31	8.69E−04	45523	43965	42408	1557
	Hierarchical model 2	6.28E+31	9.15E−06	45506	43975	42444	1531

Note: 1. The reported values are $p(\mathbf{y}|M)^*\exp(22048)$
2. The reported values are $p^*(\mathbf{y}|M)^*\exp(21736)$

Table 3: Approximated marginal densities and Bayesian deviance estimates (averaged over 10 replications) for the three MIRT models under 4 simulated situations when the additive model is true.

		$p(\mathbf{y}\|M)$	$p^*(\mathbf{y}\|M)$	DIC	$\bar{D}$	$D(\bar{\vartheta})$	p_D
Simualtion 1							
	Additive model	$2.65\text{E}{+}220^1$	$4.01\text{E}{+}195^2$	42286	40053	37819	2233
	Hierarchical model 1	$1.43\text{E}{-}116^1$	$5.02\text{E}{-}171^2$	43331	41642	39954	1689
	Hierarchical model 2	$1.44\text{E}{-}120^1$	$3.15\text{E}{-}181^2$	43314	41658	40002	1656
Simualtion 2							
	Additive model	$4.44\text{E}{+}38^3$	$4.58\text{E}{-}56^4$	39522	37619	35715	1903
	Hierarchical model 1	$1.20\text{E}{-}68^3$	$8.70\text{E}{-}196^4$	39967	38273	36579	1694
	Hierarchical model 2	$2.59\text{E}{-}85^3$	$7.5\text{E}{-}197^4$	39964	38301	36637	1663
Simualtion 3							
	Additive model	$1.38\text{E}{+}227^1$	$6.33\text{E}{+}195^2$	42331	40188	38045	2143
	Hierarchical model 1	$3.72\text{E}{-}134^1$	$1.46\text{E}{-}193^2$	43312	41800	40288	1512
	Hierarchical model 2	$8.91\text{E}{-}146^1$	$5.41\text{E}{-}195^2$	43295	41837	40380	1458
Simualtion 4							
	Additive model	$1.83\text{E}{+}155^3$	$5.48\text{E}{+}54^4$	39308	37535	35761	1773
	Hierarchical model 1	$2.33\text{E}{+}50^3$	$4.96\text{E}{-}71^4$	39688	38120	36552	1568
	Hierarchical model 2	$8.64\text{E}{+}27^3$	$3.68\text{E}{-}83^4$	39683	38120	36654	1514

Note: 1. The reported values are $p(\mathbf{y}|M)^*\exp(20440)$; 2. The reported values are $p^*(\mathbf{y}|M)^*\exp(20150)$
3. The reported values are $p(\mathbf{y}|M)^*\exp(18900)$; 4. The reported values are $p^*(\mathbf{y}|M)^*\exp(18400)$

for the additive MIRT model is no more than any of those for the two hierarchical models. However, when the additive model is true, the additive MIRT model always has a larger p_D value than the other two models.

3. When the latent ability dimensions form an additive structure, the additive MIRT

model is more superior to the hierarchical models when there are no correlations between the general and specific abilities (as shown in simulations 1 and 3), as opposed to the situation when the general and specific abilities are correlated (as shown in simulations 2 and 4).

4. Among the two hierarchical MIRT models, model 1 is more favored by the Bayes factor in all the simulated situations. However, the posterior Bayes factor and Bayesian DIC indicate that model 2 is better. Hence, there is no conclusive finding as to which of the hierarchical model performs better than the other. This is similar to the findings in Sheng and Wikle (2008).

7. An example with *CBASE* data

As an illustration, the proposed model was subsequently implemented on a subset of *CBASE English* subject data. In real test situations, the true latent structure is not necessarily known. Hence, model comparison is necessary to determine if the proposed additive MIRT model provides a relatively better representation of the data compared with other models.

7.1 Methodology

The overall *CBASE* exam contains 41 English multiple choice items, with the first 16 items forming a writing cluster and the remaining 25 a reading\literature cluster. The data used in this study were from college students who took the same form of *CBASE* in years 2001 and 2002. After removing missing responses and those who made multiple attempts, a sample of 1,231 examinees was randomly selected. To assess the goodness-of-fit, the proposed MIRT model was compared with four models, namely, the unidimensional model, the multi-unidimensional model, and the two hierarchical MIRT models. Each of the five candidate models was 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. The Gelman-Rubin R statistics were used to assess convergence and they were found to be around or close to 1, suggesting that stationarity had been reached within the simulated Monte Carlo chains for the model. Then, the five candidate models were compared using Bayes factors, Bayesian DICs and predictive model checks.

7.2 Results and Discussion

After fitting the proposed additive MIRT model to the *CBASE English* data via the Gibbs sampler, the posterior expectations of the posterior samples were used to estimate item parameters and are displayed in Table 4. The Monte Carlo (MC) standard errors of estimates are also reported in Table 4. Because subsequent samples in the Markov chain are autocorrelated, they were estimated using batching (Ripley, 1987). That is, with a long chain of samples being separated into contiguous batches of equal length, the MC

Table 4: Posterior means and Monte Carlo standard error of estimate (MCSE) for each item parameter when fitting the proposed model to the *CBASE* data.

	Posterior Mean			MCSE				Posterior Mean			MCSE		
Item	α_0	α_1	γ	α_0	α_1	γ	Item	α_0	α_1	γ	α_0	α_1	γ
1	0.3656	0.113	−0.5729	0.0055	0.0064	0.0011	21	0.5738	0.0666	−0.3271	0.0062	0.0016	0.0009
2	0.3202	0.0265	−0.6139	0.0048	0.0012	0.0014	22	0.4990	0.1154	−0.5026	0.0096	0.0024	0.0014
3	0.4027	0.0474	−1.0448	0.0047	0.0022	0.0016	23	0.5881	0.0778	−1.0336	0.0101	0.0045	0.0039
4	0.3437	0.1847	−1.4474	0.0111	0.0074	0.0012	24	0.4618	0.1834	−0.1439	0.0093	0.0021	0.0013
5	0.4135	0.1315	−1.2794	0.0086	0.0031	0.0033	25	0.3202	0.1731	−0.3153	0.0073	0.0036	0.0005
6	0.5889	0.0862	−0.8845	0.0076	0.0060	0.0020	26	0.5666	0.0347	−0.9357	0.0054	0.0012	0.0023
7	0.2169	0.0899	−0.5196	0.0059	0.0028	0.0007	27	0.2411	0.0715	−0.9282	0.0044	0.0022	0.0006
8	0.3020	0.1805	−1.228	0.0087	0.0112	0.0042	28	0.4444	0.0578	−0.6238	0.0057	0.0020	0.0004
9	0.4150	0.3997	−0.2107	0.0190	0.0123	0.0018	29	0.3391	0.3107	−0.3042	0.0117	0.0042	0.0010
10	0.5335	0.3508	−0.1145	0.0173	0.0115	0.0018	30	0.518	0.1202	−0.4452	0.0118	0.0037	0.0012
11	0.2925	0.0369	0.3662	0.0046	0.0012	0.0008	31	0.4053	0.3954	−0.8646	0.0156	0.0057	0.0010
12	0.3356	0.0784	−0.7815	0.0052	0.0057	0.0014	32	0.5058	0.3543	−1.077	0.0143	0.0079	0.0037
13	0.4114	0.0471	−0.1872	0.0063	0.0024	0.0014	33	0.2446	0.1699	−0.4488	0.0052	0.0044	0.0008
14	0.3001	0.2491	−0.1768	0.0096	0.0069	0.0012	34	0.2389	0.4873	−0.8346	0.0150	0.0076	0.0028
15	0.5562	0.1476	−0.8749	0.0065	0.0109	0.0037	35	0.3172	0.3600	−0.2555	0.0133	0.0068	0.0010
16	0.3415	0.2158	−0.1082	0.0099	0.0051	0.0012	36	0.3236	0.1766	0.3571	0.0078	0.0016	0.0003
							37	0.2986	0.3209	0.3177	0.0131	0.0053	0.0017
17	0.4030	0.0345	−0.158	0.0040	0.0006	0.0005	38	0.2873	0.2522	−0.5023	0.0102	0.0024	0.0012
18	0.3411	0.0672	−0.3315	0.0042	0.0020	0.0011	39	0.4437	0.3707	−0.7481	0.0146	0.0068	0.0015
19	0.5785	0.1516	0.3097	0.0090	0.0031	0.0009	40	0.2761	0.5462	−0.4558	0.0184	0.0098	0.0023
20	0.8620	0.0695	−1.4365	0.0187	0.0043	0.0072	41	0.1674	0.2363	−0.3417	0.0080	0.0033	0.0003

standard error for each parameter is estimated to be the standard deviation of these batch means, and the MC standard error of estimate is then a ratio of the standard deviation and the square root of the number of batches. Generally, all the standard errors for the posterior estimates of the item parameters were small, with those for item difficulties, γ, being relatively smaller. It can be interpreted that, for example, an approximate 99% MC interval for the true posterior expectation for the first item's discrimination parameter associated with the general ability was $0.3656 \pm 3 \times (0.0055)$, suggesting the MC estimate of this posterior mean was good to about two digits of accuracy. Hence, the item parameters using the proposed Bayesian models were estimated with little error.

The model choice measures were subsequently obtained and the results are summarized as follows. Table 5 displays the results for Bayes factors and Bayesian deviances. The first two columns are the approximated marginal densities $p(\mathbf{y}|M)$ and $p^*(\mathbf{y}|M)$ for the five candidate models. As a *BF* or *PBF* greater than 100 indicates decisive evidence in favor of the model on the numerator, the additive MIRT model was found to be the best among the five candidate models. Taking the ratio of its marginal density with that for any other models resulted in *BF* or *PBF* estimates greater than 100. On the other hand, there is much evidence against the unidimensional model when comparing it to either the multi-unidimensional model, the hierarchical MIRT models or the proposed additive MIRT model. Moreover, the hierarchical MIRT model 1 was shown to be better than the multi-unidimensional model using the *BF*, but not the *PBF* estimate.

Table 5 also displays the Bayesian deviance results, where smaller values indicate better model fit. Among the five candidate IRT models, the proposed additive MIRT model had the smallest DIC and expected posterior deviance ($\bar{D}$). Therefore, the additive MIRT

Table 5: Approximated marginal densities of the data and Bayesian deviance estimates for the five IRT models with the *CBASE* data.

	$p(\mathbf{y}\|M)^1$	$p^*(\mathbf{y}\|M)^2$	DIC	$\bar{D}$	$D(\bar{\vartheta})$	p_D
Unidimensional	1.2254E−224	8.55E−308	55639	54548	53457	1090.6
Multi-unidimensional	4.2856E−163	1.04E−207	55571	54160	52750	1410.5
Hierarchical model 1	2.568E−143	2.83E−215	55586	54121	52656	1464.6
Hierarchical model 2	8.0348E−177	4.6805E−220	55571	54188	52805	1382.7
Additive model	156	107.5633	55135	53318	51501	1817.3

Note: 1. The reported values are $p(\mathbf{y}|M)^*\exp(26840)$
2. The reported values are $p^*(\mathbf{y}|M)^*\exp(26460)^*4000$

model provided the best goodness-of-fit to the data compared with other models, even after penalizing for a large effective number of parameters ($p_D = 1817.3$), which is shown in the last column of the table. Compared with the multi-unidimensional model, the two hierarchical MIRT models did not show much better fit to the data using Bayesian DICs. In addition, the additive MIRT model had a larger p_D than the two hierarchical models. Given the findings from the simulation study in Section 6, this indicated that the latent structure for the general and specific abilities was closer to additive. On the other hand, the unidimensional model was relatively worse than any of the multidimensional models. The results were generally consistent with those obtained using the Bayes factors.

Next, the posterior predictive model checking procedure was implemented to compare the five candidate models. To do so, a test statistic had to be chosen for describing the discrepancy between the model and the data. For this analysis, the odds ratio was adopted for measuring association among item pairs, $\mathrm{T}(\mathbf{y}) = \mathrm{OR}_{ij} = \dfrac{n_{11}n_{00}}{n_{01}n_{10}}$,where $n_{kk'}$ denotes the number of examinees scoring k on item i and k' on item j, $k, k' = 0, 1$. This statistic has been reported to be powerful for detecting unidimensionality in data (Sinharay et al., 2006). Hence, for each fitted model, based on each pair of $(\boldsymbol{\theta}, \boldsymbol{\xi})$ samples, a $\mathbf{y}^{\mathrm{rep}}$ was simulated and the replicated odds ratios $\mathrm{T}(\mathbf{y}^{\mathrm{rep}})$ were computed and further compared with the actual odds ratios. The tail-area PPP-values (p_B) were estimated as the proportion of the simulated samples for which $\mathrm{T}(\mathbf{y}^{\mathrm{rep}}) \geq \mathrm{T}(\mathbf{y})$, i.e., $p_B = \sum_{l=1}^{L} I(\mathrm{T}(\mathbf{y}^{\mathrm{rep}l}) \geq \mathrm{T}(\mathbf{y}))$.

Figure 2 summarizes the extreme PPP-values for the odds ratios with each model. Here $\alpha = .05$ was used as the critical level, so that the PPP-value larger than .975 was denoted using a plus sign and the PPP-value smaller than .025 was denoted using a cross sign. Since odds ratios are based on the responses to any pair of items, each plot is symmetrical about its diagonal. Hence, the upper-diagonal was left blank for simplicity. From the figure, it is immediately clear that the proposed additive MIRT model had far fewer extreme replicated odds ratios. Indeed, the numbers of extreme PPP-values for the five candidate models, namely, the unidimensional, multi-unidimensional, two hierarchical MIRT, and the proposed additive MIRT models, were 39, 36, 37, 37 and 12, respectively. The additive model showed remarkably less error in predicting odds ratios for item pairs within clusters as well as those between clusters and is considered to be the best among the five candidate models. On the other hand, the unidimensional IRT model had the

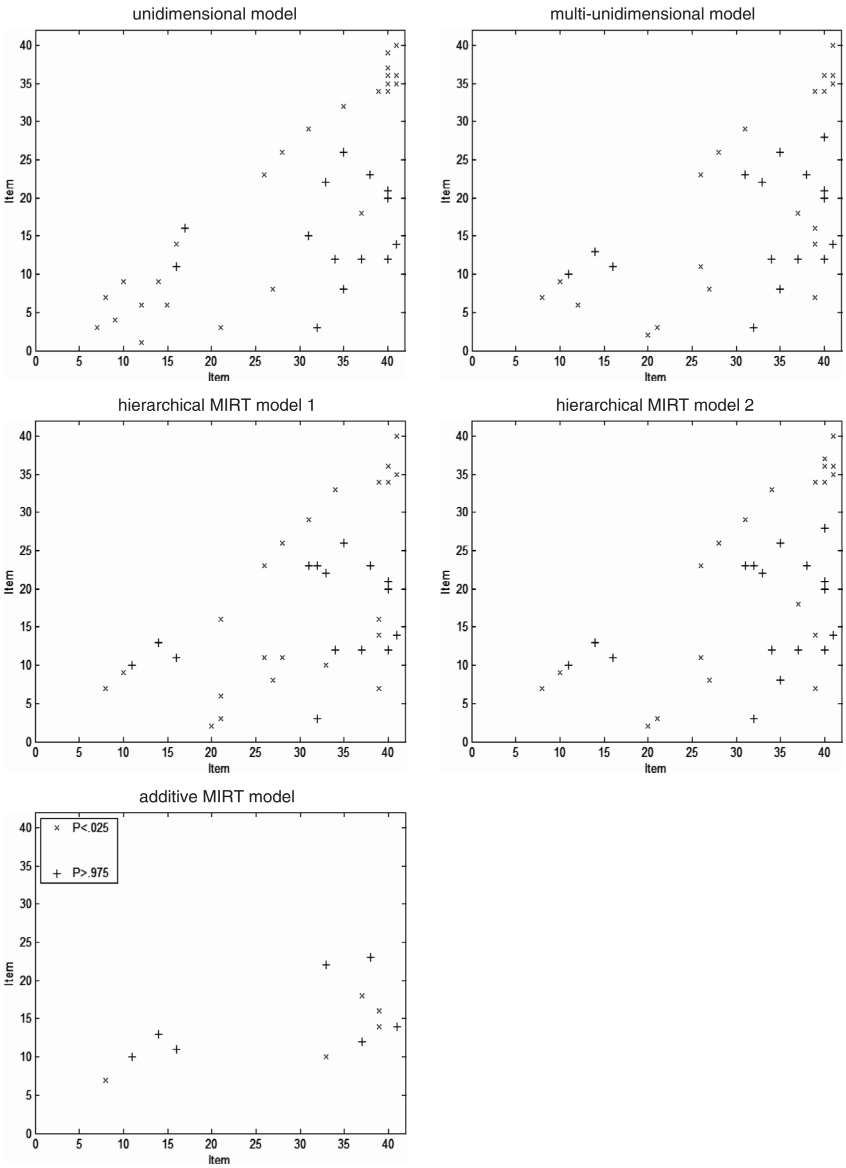


Figure 2: Extreme tail-area PPP-values for odds ratios with the five IRT models for the *CBASE* data.

largest number of extreme PPP-values and hence is shown to be the worst using the odds ratio for posterior model checks. Although with slightly different prediction errors, the two hierarchical MIRT models performed similarly in their abilities to predict the odds ratio, which were not much different from the multi-unidimensional model.

Therefore, with Bayesian model checking techniques, the five candidate IRT models were evaluated as to which model provided a better description, and hence a better goodness-of-fit to the *CBASE* data. The results from Bayes factors, Bayesian deviances and posterior predictive checks all showed strong evidence in favor of the proposed additive model, which was believed to fit the data much better than the other candidate models. On the contrary, 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.

8. Discussion and Conclusion

In conclusion, IRT-based models incorporating both general ability and specific abilities so that they directly affect how examinees answer each test item can be developed from several perspectives. As the proposed model specifies a generalized linear function of the general ability and a specific ability, the multicollinearity problem associated with the linear models might potentially affect the accuracy of parameter estimation. Hence, the additive MIRT model performs relatively better when the general ability and each specific ability are less highly related. This is shown to be the case from the simulation studies. In addition, the proposed additive MIRT model, using an MCMC procedure, performs consistently better than the hierarchical MIRT models in various simulated test situations and even when the latent structure of the general and specific abilities is not additive. However, when the latent structure is additive, the additive MIRT model tends to have a larger effective number of parameters than the two hierarchical MIRT models. This may serve as an indicator on the actual latent structure with real data. Furthermore, the proposed additive MIRT model is implemented on the *CBASE English* data via Gibbs sampling with small standard errors. This suggests both general ability and specific ability dimensions can be estimated in one implementation with enough accuracy. As far as the *CBASE* data is concerned, the proposed model provides a better description to the data than the conventional unidimensional model, the multi-unidimensional model, or the two hierarchical MIRT models. Consequently, the proposed additive MIRT model offers a better way to represent the test situations not realized in existing models.

To paraphrase Box (1976), it is well accepted that all theoretical models are just simplified approximations of the real world. Some models represent reality better than others. Therefore, it is vitally important to find the model(s) providing the most complete description of the data. In testing situations where IRT models are used for parameter estimation as well as other applications, one has to decide the dimensionality structure for the latent abilities in order to choose an appropriate model and hence obtain reliable estimates of person abilities. Often, a unidimensional model is adopted by assuming one latent ability. However, this assumption is more likely to be violated in real situations because the test items are not always measuring a single trait. This point is easily seen from the findings of the current study, where model comparisons indicate that the unidimensional model describes the *CBASE* data the worst compared with models with multiple dimensions. Therefore, using the unidimensional model for the *CBASE English* test is not validated. The actual dimensionality for the test is closer to the structure with one general and two specific ability dimensions so that they form an additive structure. In particular, the first 16 test items measure the overall English ability and a writing ability, and the last 25 items measure the overall ability together with a reading/literature ability. All items are affected by a general ability and a specific ability simultaneously and directly. However, the actual relationship between the general ability and each of the two specific ability dimensions cannot be estimated directly given the limitation of the model specification noted in Section 3. Further studies are needed for a better solution.

In the current study, odds ratios were adopted as a discrepancy measure for the predictive model checking technique. Other test statistics could also be considered, such as item test biserial correlations, observed score distribution, or test information function, among others. The choice of discrepancy measures is crucial with the method, as some measures may fail to detect the differences between models, such as item proportion-correct (Sinharay et al., 2006). However, we note that this procedure has been criticized for being conservative and the PPP-value is not uniformly distributed under the null hypothesis (Sinharay & Stern, 2003). Future studies can adopt other methods for comparing models, such as looking at the Bayesian residuals as proposed by Albert and Chib (1995). Additionally, in our study, Bayes factors 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 a "rare" value with small likelihood, which has a large effect on the final result. Future studies can adopt more accurate methods that are based on estimation of marginal likelihoods, such as the Chib's method (Chib, 1995; Chib & Jeliazkov, 2001) or the bridge sampling method (Meng & Wong, 1996; Meng & Shiling, 2002). Finally, in the proposed model, a strong prior was adopted for the ability parameters by using the identity matrix as the covariance matrix to avoid the model indeterminacy problem. Future study may employ other approaches to resolve this nonidentifiability problem.

Appendix A. Comparing the additive model with the testlet model

A simulation study was carried out to compare the additive model with the more specific testlet model. In the study, four simulations were conducted, where tests with one general ability and two specific abilities were considered, i.e., $m = 2$. Dichotomous item responses of 1,000 persons to 41 items were simulated so that the responses y_{vij} were generated from the probabilities as defined in (3), where $\alpha_{0vj} \sim U(0,1)$, $\alpha_{vj} \sim U(0,1)$, $\gamma_{vj} \sim U(-1,1)$. In addition, the ability parameters $\boldsymbol{\theta}_i$ were simulated from $N_3(\mathbf{0}, \mathbf{R}_0)$, where $\mathbf{R}_0$ is a correlation matrix and was specified to be $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0.8 & 1 & \\ 0.6 & 0 & 1 \end{pmatrix}$, $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0.6 & 1 \end{pmatrix}$, and $\mathbf{R}_0 = \begin{pmatrix} 1 & & \\ 0.8 & 1 & \\ 0.6 & 0.5 & 1 \end{pmatrix}$ in the four simulations, respectively.

With the simulated responses, the additive and the testlet models were each imple-

Table A1: Approximated marginal densities and Bayesian deviance estimates (averaged over 10 replications) for the additive model and the testlet model under 4 simulated situations.

		$p(\mathbf{y}\|M)$	$p^*(\mathbf{y}\|M)$	DIC	$\bar{D}$	$D(\bar{\vartheta})$	p_D
Simualtion 1							
	Additive model	3.21E+144^1	7.57E+261^1	42491	40265	38039	2226
	Testlet model	2.03E−163^1	2.46E−66^1	43494	41812	40130	1682
Simualtion 2							
	Additive model	4.10E+78^2	6.23E+194^2	39400	37484	35569	1916
	Testlet model	3.23E−46^2	3.45E+43^2	39883	38124	36365	1759
Simualtion 3							
	Additive model	4.08E+175^1	1.69E+294^1	42257	40126	37996	2131
	Testlet model	9.51E−133^1	1.85E−40^1	43322	41668	40013	1654
Simualtion 4							
	Additive model	1.07E+153^2	5.35E+278^2	39203	37431	35660	1771
	Testlet model	1.27E+50^2	6.20E+135^2	39693	37950	36207	1743

Note: 1. The reported values are $p(\mathbf{y}|M)^*\exp(20500)$ or $p^*(\mathbf{y}|M)^*\exp(20500)$
3. The reported values are $p(\mathbf{y}|M)^*\exp(18800)$ or $p^*(\mathbf{y}|M)^*\exp(18800)$

mented using Gibbs sampling where 7,000 iterations were obtained with the first 2,000 as burn-in, which was sufficient for the chains to reach stationarity. Ten replications were used and the posterior expectations of the Gibbs samples were used to obtain the posterior estimates necessary to derive Bayes factors as well as Bayesian deviance (see Section 4 for a description of these measures) results.

The model comparison results in each simulation were averaged over the ten replications and are summarized in Table A1. The marginal densities $p(\mathbf{y}|M)$ and $p^*(\mathbf{y}|M)$, displayed in the first two columns of the table, are used to compute the Bayes factor (BF) and the posterior Bayes factor (PBF) between two models M_i and M_j, i.e., $BF = \dfrac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)}$, $PBF = \dfrac{p^*(\mathbf{y}|M_i)}{p^*(\mathbf{y}|M_j)}$. As a BF or PBF greater than 100 indicates decisive evidence in favor of M_i (cf., Robert, 2001), the additive MIRT model was found to be consistently better than the more strict testlet model, even when the actual intertrait correlations were zero (because the testlet model assumes that α_{0vj} and α_{vj} are equal whereas they were set differently in the simulation study).

The remaining table summarizes the Bayesian deviance results. Specifically, the additive MIRT model shows consistently smaller DIC, $\bar{D}$, and $D(\bar{\vartheta})$, than the testlet model. Since small deviance values indicate better model fit, the additive MIRT model is suggested to provide a better description of the simulated data in various simulated situations considered, even after penalizing for model complexities. Hence, the testlet model was not considered in the analysis of the study.

Appendix B. Full conditional distributions for the Bayesian additive MIRT model

The full conditional distribution for each parameter can be derived as follows:

1. For variable Z_{vij}:

$$\begin{aligned}[Z_{vij}|\bullet] &\propto f(y_{vij}|Z_{vij})p(Z_{vij}|\eta_{vij}) \\ &\propto \exp\{-\frac{1}{2}(Z_{vij}-\eta_{vij})^2\}(I(Z_{vij}>0)I(y_{vij}=1)+I(Z_{vij}\le 0)I(y_{vij}=0)). \end{aligned} \tag{8}$$

So, the full conditional of Z_{vij}, denoted as $Z_{vij}|\bullet$ has as a truncated normal distribution

$$Z_{vij}|\bullet \sim \left\{ \begin{array}{lll} N_{(0,\infty)}(\eta_{vij},1), & if & y_{vij}=1 \\ N_{(-\infty,0)}(\eta_{vij},1), & if & y_{vij}=0 \end{array} \right. . \tag{9}$$

2. For the person parameters $\boldsymbol{\theta}_i$:

$$\begin{aligned} [\boldsymbol{\theta}_i|\bullet] &\propto p(\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\xi})p(\boldsymbol{\theta}|\boldsymbol{\mu}) \\ &\propto \exp\{-\frac{1}{2}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_i)'(\boldsymbol{\theta}_i-\boldsymbol{\mu}_i)\}\prod_{v=1}^{m}\prod_{j=1}^{k_v}\exp\{-\frac{1}{2}(Z_{vij}-(\alpha_{0vj}\theta_{0i}+\alpha_{vj}\theta_{vi}-\gamma_{vj}))^2\} \\ &= \exp\{-\frac{1}{2}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_i)'(\boldsymbol{\theta}_i-\boldsymbol{\mu}_i)\}\exp\{-\frac{1}{2}(\boldsymbol{A}\boldsymbol{\theta}_i-\mathbf{B})'(\boldsymbol{A}\boldsymbol{\theta}_i-\mathbf{B})\} \\ &\propto \exp\{-\frac{1}{2}[\boldsymbol{\theta}_i'(\mathbf{A}'\mathbf{A}+\mathbf{I})\boldsymbol{\theta}_i-2(\boldsymbol{\mu}_i+\mathbf{A}'\mathbf{B})'\boldsymbol{\theta}_i]\}. \end{aligned} \tag{10}$$

Thus, the full conditional for $\boldsymbol{\theta}_i$ has a multivariate normal distribution,

$$\boldsymbol{\theta}_i|\bullet \sim N_{m+1}((\mathbf{A}'\mathbf{A}+\mathbf{I})^{-1}(\boldsymbol{\mu}_i+\mathbf{A}'\mathbf{B}),(\mathbf{A}'\mathbf{A}+\mathbf{I})^{-1}), \tag{11}$$

where $\mathbf{A}=\begin{pmatrix} \alpha_{01} & \alpha_1 & 0 & \cdots & 0 \\ \alpha_{02} & 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{0m} & 0 & 0 & \cdots & \alpha_m \end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix} Z_{1i}+\gamma_1 \\ Z_{2i}+\gamma_2 \\ \vdots \\ Z_{mi}+\gamma_m \end{pmatrix}$.

3. Then, for the item parameters $\boldsymbol{\xi}_{vj}$:

$$\begin{aligned} [\boldsymbol{\xi}_{vj}|\bullet] &\propto p(\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\xi})p(\boldsymbol{\xi}) \\ &\propto \prod_{i=1}^{n}\exp\{-\frac{1}{2}(Z_{vij}-(\alpha_{0vj}\theta_{vi}+\alpha_{vj}\theta_{vi}-\gamma_{vj}))^2\}\exp\{-\frac{1}{2}\boldsymbol{\xi}_{vj}'\boldsymbol{\xi}_{vj}\}I(\alpha_{0vj}>0)I(\alpha_{vj}>0) \\ &= \exp\{-\frac{1}{2}[(\mathbf{Z}_{\mathrm{v}}-\mathbf{x}_v\boldsymbol{\xi}_{vj})'(\mathbf{Z}_v-\mathbf{x}_v\boldsymbol{\xi}_{vj})+\boldsymbol{\xi}_{vj}'\boldsymbol{\xi}_{vj}]\}I(\alpha_{0vj}>0)I(\alpha_{vj}>0) \\ &\propto \exp\{-\frac{1}{2}[\boldsymbol{\xi}_{vj}'(\mathbf{x}_v'\mathbf{x}_v+\mathbf{I})\boldsymbol{\xi}_{vj}-2(\mathbf{Z}_v'\mathbf{x}_v)\boldsymbol{\xi}_{vj}]\}I(\alpha_{0vj}>0)I(\alpha_{vj}>0). \end{aligned} \tag{12}$$

So, the full conditional for $\boldsymbol{\xi}_{vj}$ is

$$\boldsymbol{\xi}_{vj}|\bullet \sim N((\mathbf{x}_v'\mathbf{x}_v + \mathbf{I})^{-1}\mathbf{x}_v'\mathbf{Z}_v, (\mathbf{x}_v'\mathbf{x}_v + \mathbf{I})^{-1})I(\alpha_{0vj} > 0)I(\alpha_{vj} > 0), \tag{13}$$

where, $\mathbf{Z}_v = [Z_{vij}]_{nxk_v}$, $\boldsymbol{\xi}_v = (\boldsymbol{\xi}_{v1}, \ldots, \boldsymbol{\xi}_{vk_v})'$, $\mathbf{x}_v = [\boldsymbol{\theta}_0, \boldsymbol{\theta}_v, -1]$, and $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0n})'$, $\boldsymbol{\theta}_v = (\theta_{v1}, \ldots, \theta_{vn})'$, $v = 1, \ldots, m$.

4. Next, for the hyperparameter $\boldsymbol{\mu}_i$:

$$\begin{aligned}[\boldsymbol{\mu}_i|\bullet] &\propto p(\boldsymbol{\theta}|\boldsymbol{\mu})p(\boldsymbol{\mu}) \propto \exp\{-\tfrac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)'(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)\}\exp\{-\tfrac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\} \\ &\propto \exp\{-\tfrac{1}{2}[\boldsymbol{\mu}_i'(\mathbf{I} + \boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}_i - 2\boldsymbol{\theta}_i'\boldsymbol{\mu}_i]\}.\end{aligned} \tag{14}$$

So, the full conditional for $\boldsymbol{\mu}_i$ is distributed as

$$\boldsymbol{\mu}_i|\bullet \sim N_{m+1}((\mathbf{I} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\theta}_i, (\mathbf{I} + \boldsymbol{\Sigma}^{-1})^{-1}). \tag{15}$$

5. Lastly, for the hyperparameter $\boldsymbol{\Sigma}$:

$$\begin{aligned}[\boldsymbol{\Sigma}|\bullet] &\propto p(\boldsymbol{\mu}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{2(m+1)+1}{2}}\exp\{-\tfrac{1}{2}tr(\boldsymbol{\Sigma}^{-1})\}\prod_{i=1}^{n}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\} \\ &= |\boldsymbol{\Sigma}|^{-\frac{2(m+1)+n+1}{2}}\exp\{-\tfrac{1}{2}tr[(\mathbf{S} + \mathbf{I})\boldsymbol{\Sigma}^{-1}])\}.\end{aligned} \tag{16}$$

Thus, the full conditional for $\boldsymbol{\Sigma}$ is an inverse Wishart distribution,

$$\boldsymbol{\Sigma}|\bullet \sim W^{-1}((\mathbf{S} + \mathbf{I})^{-1}, n + m + 1), \tag{17}$$

where $\mathbf{S} = \sum_{i=1}^{n}\boldsymbol{\mu}_i\boldsymbol{\mu}_i'$

References

Ackerman, T.A. (1993). *Insuring the validity of the reported score scale by reporting multiple scores*. Paper presented at the North American Meeting of the Psychometric Society, Berkeley, CA.

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 111–142.

Albert, J.H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, **17**, 251–269.

Albert, J.H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, **66**, 541–562.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, **64**, 153–168.

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

Carlin, B.P., & Chib, S. (1993). *Bayesian model choice via Markov chain Monte Carlo*. Research Report 93–006. University of Minnesota, Division of Biostatistics.

Carlin, B.P., & Louis, T.A. (2000). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**, 1197–1208.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis.* Boca Raton: Chapman & Hall/CRC.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Lee, H. (1995). *Markov chain Monte Carlo methods for estimating multidimensional ability in item response analysis.* Ph.D. Dissertation, University of Missouri, Columbia, MO.

Meng, X.L., & Wong, W.H. (1996). Simulating ratios of normaliing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, **6**, 831–860.

Meng, X.L., & Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, **11**, 552–586.

Newton, M.A., & Raftery, A.E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B*, **56**, 3–48.

Osterlind, S. (1997). *A national review of scholastic achievement in general education: How are we doing and why should we care?* ASHE-ERIC Higher Education Report 25, No. 8. Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, **21**, 25–36.

Ripley, B.D. (1987). *Stochastic Simulation.* New York: Wiley.

Robert, C.P. (2001). *The Bayesian Choice* (2$^{\text{nd}}$ ed). New York: Springer.

Schmid, J., & Leiman, J.M. (1957). The development of hierarchical factor solutions. *Psychometrika*, **22**, 53–61.

Segall, D.O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, **66**, 79–97.

Sheng, Y., & Wikle, C.K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational & Psychological Measurement*, **67**, 899–919.

Sheng, Y., & Wikle, C.K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational & Psychological Measurement*, **68**, 413–430.

Sinharay, S., Johnson, M.S., and Stern, H.S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, **30**, 298–321.

Sinharay, S., & Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, **111**, 209–221.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American journal of Psychology*, **15**, 201–293.

Spiegelhalter, D.J., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, *Series B*, **64**, 583–640.

Tanner, M.A., and W.H. Wong (1987). The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association,* **82**, 528–550.

(Received October 24 2007, Revised November 10 2008)