Comparing Multiunidimensional and Unidimensional Item Response Theory Models
Yanyan Sheng and Christopher K. Wikle

Educational and Psychological Measurement 2007 67: 899 originally published online 25 May 2007
DOI: 10.1177/0013164406296977

The online version of this article can be found at:
http://epm.sagepub.com/content/67/6/899

Published by:
SAGE
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: http://epm.sagepub.com/content/67/6/899.refs.html

>> Version of Record - Nov 21, 2007
Proof - May 25, 2007
What is This?
Comparing Multiunidimensional and Unidimensional Item Response Theory Models

Yanyan Sheng
Southern Illinois University Carbondale
Christopher K. Wikle
University of Missouri-Columbia

For tests consisting of multiple subtests, unidimensional item response theory (IRT) models apply when the subtests are known to measure a common underlying ability. However, in many instances, due to the lack of a satisfactory index for assessing the dimensionality assumption, the test structure is not clear. A more general IRT model, the multiunidimensional model, is more flexible and efficient in various test situations. This article compares these two classes of normal ogive two-parameter models and shows that the multiunidimensional model offers a better way to represent test situations not realized in unidimensional models.

Keywords: item response theory; unidimensional model; multiunidimensional model; Markov chain Monte Carlo; Bayesian model choice

Item response theory (IRT; Lord, 1980) has gained increasing popularity in large-scale educational and psychological testing situations. Modeling the interaction of a person’s ability and the test at the item level, IRT models have advantages over traditional classical test theory models (Hambleton, Swaminathan, & Rogers, 1991). These advantages are apparent in numerous practical applications such as ability estimation, equating, differential item functioning, and computerized adaptive testing.

In many circumstances, it suffices to assume that all the test items measure one ability in common and hence use unidimensional IRT models. However, in other situations when it is a priori clear that multiple abilities are being measured or the test dimensionality structure is not clear at all, multidimensional IRT (MIRT) models have to be considered. Much work has been conducted to develop multidimensional models for dichotomous scored items (e.g., Béguin & Glas, 2001; Bock & Aitkin, 1981; Hoijtink & Molenaar, 1997). The present study focuses on a special

Authors’ Note: Please address correspondence to Yanyan Sheng, Department of Educational Psychology & Special Education, Mailcode 4618, Southern Illinois University, Carbondale, IL 62901; e-mail: ysheng@siu.edu.
type of multidimensional model and illustrates its efficiency compared to the unidimensional model under various conditions.

Under IRT, an examinee’s response to a specific test item or question is determined by an unobserved mental trait of the examinee. Each of these underlying traits, or what are usually called abilities, is assumed to vary continuously along a single dimension, denoted by $\theta$, so that the examinees responding to the test items can be arrayed on $\theta$, from lowest to highest. The position of person $i$ on $\theta$, denoted $\theta_i$, is usually referred to as the person’s ability or proficiency. Intuitively, we expect the probability of a correct response to each item to increase monotonically as $\theta_i$ increases. In terms of binary scored test items, that is, items on which responses are designated either correct or incorrect, the IRT models express the probability of a correct response to a test item as a function of $\theta$ given one or more parameters of the item. These models are generally referred to as unidimensional models and are appropriate in situations where all test items are designed to measure one ability dimension, $\theta$, in common.

Unidimensional IRT models rely on a strong assumption that each test item is designed to measure some facet of the same underlying ability or so-called unified latent trait. It is necessary that a test intending to measure one certain trait should not be affected by other traits, especially when only the overall test scores are reported and used as an assessment criterion for various ability levels. However, Hattie (1985) reviewed various techniques for checking unidimensionality and found none of them fully satisfactory in different conditions of a test. Likewise, many Monte Carlo studies conducted to evaluate various dimensionality indexes have advised caution when generalizing results to other conditions (test length, sample size, etc.; De Champlain & Gessaroli, 1998; Gessaroli & De Champlain, 1996; Hattie, Krakowski, Rogers, & Swaminathan, 1996). Research has shown that when a test known to be multidimensional is modeled using a unidimensional model, the parameter estimation will be biased (Folk & Green, 1989). In addition, measurement error increases and incorrect inferences about an examinee’s proficiency in a given subject may be made (Walker & Beretvas, 2000). Therefore, the validity of useful, meaningful, and appropriate interpretations of the test scores is questionable. Indeed, as McNemar (1946) states, regarding the composite of two types of items, “unless highly correlated, the meaning of scores based on such a composite is questionable” (p. 298).

MIRT models are adopted when distinct multiple abilities (i.e., more than one $\theta$) are involved in producing the manifest responses for an item. MIRT is closely related to factor analysis despite the different focuses of the two approaches (Reckase, 1997). More specifically, the unidimensional IRT model is appropriate when only one factor is extracted from the test items, whereas MIRT models are adopted when more than one factor is found to be significant. In the latter case, it should be noted that each item measures multiple abilities, that is, each test item has loadings on all factors extracted. In an exploratory factor analysis, the factor solution is usually visually or
analytically rotated. Often the rotation scheme is devised to approximate simple structure (McDonald, 1985) so that the factor loadings are split into two groups: the elements of one tending to zero and those of the other tending toward unity. Hence, each item has a unity loading on one factor and zero loadings on other factors. Analogously, in the MIRT setting, the test involves multiple abilities, and each item measures only one of them. This is probably more common in large-scale testing situations than the more general case where each item measures multiple traits as described earlier. More often, a test consists of several subtests with each focusing on one specific ability and the items in a particular subtest designed to measure one ability in common. The IRT models that are appropriate in this situation are our major focus in this study. In the literature, they are called MIRT models with simple structure or between-items MIRT models. However, we prefer to use the term multiunidimensional model because it is shorter and more intuitive, that is, the test involves multiple dimensions, and each test is unidimensional by itself.

In many test situations, the actual structure of the ability dimensions is not known in advance. Hence, it is hard to decide which model to adopt. Suppose an English test consists of three subtests: listening, reading, and writing. Intuitively, the three subtests are related. One can assume all items measure a unified English ability dimension and fit a unidimensional model. However, the estimates might be biased if the subtests are not exactly measuring one latent trait. Alternatively, one may choose to fit the unidimensional model separately for each subtest to estimate the examinee’s abilities in each subdimension. However, the subscores obtained this way apply only when there are no correlations between the subtests because separate implementations of the unidimensional model do not account for the relationship between different ability dimensions. In this sense, the two approaches using the unidimensional model are restricted. It is then more appropriate to use the multiunidimensional model, which yields more precise estimates and hence is more efficient because strength can be drawn from responses of correlated ability dimensions (Lee, 1995). In model estimation, fully Bayesian methodology has been adopted and found appealing for unidimensional IRT models (e.g., Albert, 1992; Patz & Junker, 1999) as well as MIRT models (e.g., Béguin & Glas, 2001; Hoijtink & Molenaar, 1997; Segall, 2002). However, to date, no study has yet focused on Bayesian multiunidimensional models and illustrated their effectiveness.

The current study, therefore, uses Bayesian methods to illustrate the flexibility and efficiency of the normal ogive multiunidimensional model by comparing it with the usual unidimensional model. That is, we use formal Bayesian model selection to compare the unidimensional and multiunidimensional models. The advantages of the approach are twofold. First, this fully Bayesian approach incorporates the dependencies among variables and sources of uncertainty. Second, the computations for estimating parameters are based on Monte Carlo methods, which are free from the limitations of using Gaussian quadrature in the marginal maximum likelihood framework (Béguin & Glas, 2001).
The remainder of this article is organized as follows. First, we outline IRT models and describe the hierarchical parameterizations adopted for the unidimensional and multiunidimensional models. Next, the Bayesian model selection techniques are illustrated. Simulation studies are conducted to assess the two types of IRT models and their performance in recovering parameters under different test situations, and the results are summarized. Then, we give an example wherein the hierarchical models are implemented on a subset of College Basic Academic Subjects Examination (CBASE) English subject data, and Bayesian model selection procedures are subsequently performed for model comparisons. Finally, a few summary remarks are given.

**IRT Models**

Parametric models in the IRT literature include the Rasch (or one-parameter, 1P) model, the two-parameter (2P) model, and the three-parameter (3P) model. The Rasch model assumes items differ only in difficulty levels, whereas in the 2P model, items differentiate examinees differently. The 3P model, however, is distinguished by an additional factor that accounts for guessing. Furthermore, because these models consider binary responses, the probability functions usually take the form of a probit or logit function. As is well known, the logistic model is nearly identical to the normal ogive model (Birnbaum, 1968). In this article, we focus on mainly the 2P normal ogive (probit) models.

**Multidimensional 2P Normal Ogive IRT Model**

MIRT is a methodology that considers in psychological measurement when multiple latent traits affect the examinee’s performance on a given item (Reckase, 1997). MIRT models allow separate inferences to be made about an examinee for each distinct dimension being measured by introducing person ability and item discrimination parameters for each skill being measured by a test question (Ackerman, 1993). Suppose a test consists of \( k \) multiple choice items, each measuring \( m \) ability dimensions, \( \theta_i, \ldots, \theta_m \). Let \( y = [y_{ij}]_{n \times k} \) represent a matrix of \( n \) examinees’ responses to \( k \) dichotomous items, so that \( y_{ij} \) is defined as

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

Then, the 2P normal ogive MIRT model is defined as

\[
P(y_{ij} = 1| \theta_i, \alpha_j, \gamma_j) = \Phi \left( \sum_{v=1}^{m} \alpha_{vj} \theta_{vi} - \gamma_j \right) = \int_{-\infty}^{\frac{\theta - \gamma_j}{\sqrt{2} \sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, \quad (1)
\]
where $\theta_i = (\theta_{i1}, \ldots, \theta_{im})'$ is a vector of $m$ ability parameters and $\alpha_j = (\alpha_{1j}, \ldots, \alpha_{mj})$ is a vector of discrimination parameters for an item measuring multiple dimensions. The larger a particular $\alpha_{vj}$ is, the more important that dimension becomes in determining an examinee’s success on item $j$. In addition, $\gamma_j$ is a scalar parameter determining the location in the latent space where the item is maximally informative. It is related to the difficulty parameter in the unidimensional IRT model through a function that includes the vector of discrimination parameters and $\gamma_j$. Due to the additive nature of the ability parameters, an examinee with a low ability on one dimension is able to compensate by having a higher ability on another, thus increasing his or her probability of correctly responding to the item. Hence, the model is known as a compensatory model, in contrast to the noncompensatory MIRT model (for more illustration and comparison, see, e.g., Ackerman, 1989; Reckase, 1997). A greater estimation challenge associated with the latter makes it less practical.

**Multiunidimensional 2P Normal Ogive IRT Model**

An item’s vector of discrimination parameters $\alpha_j = (\alpha_{1j}, \ldots, \alpha_{mj})$, as specified in the MIRT model in Equation 1, can be interpreted analogous to factor loadings in factor analysis. If a rotation is performed so that each item loads on one factor only, the vector of discrimination parameters can be simplified to $\alpha_j = (0, \ldots, 0, \alpha_{vj}, 0, \ldots, 0)$, and the model is then reduced to what we call the multiunidimensional model. Consider a $k$-item test consisting of $m$ subtests, each containing $k$ multiple choice items that measure one ability dimension. With a probit link, the probability of person $i$ obtaining a correct response for item $j$ of the $v$th subtest can be defined as follows:

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

(Lee, 1995), where $\alpha_{vj}$ and $\theta_{vj}$ are scalar parameters representing the item discrimination and the examinee ability in the $v$th ability dimension, and $\gamma_{vj}$ is a scalar parameter indicating the location in that dimension where the item provides maximum information. When each item loads on one factor only, the model can be viewed as a special case of the MIRT models.

**Unidimensional 2P Normal Ogive IRT Model**

The unidimensional IRT model is a subset of the multiunidimensional model when the $m$ ability dimensions that the test measures are in effect the same. Suppose a $k$-item
test is measuring a single unified ability. Then the probability of person \(i\) obtaining a correct response for item \(j\) can be defined as follows:

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

where \(\alpha_j\) is a scalar parameter describing the item discrimination, \(\gamma_j\) is associated with item difficulty \(b_j\) such that \(\gamma_j = \alpha_j b_j\), and \(\theta_i\) is a scalar ability parameter. Among the three classes of IRT models discussed here, the unidimensional model is the most specific; thus, it requires the most stringent assumptions.

Most large-scale standardized tests are those with several subtests, each measuring a slightly different ability dimension, so a multiunidimensional IRT model is more appropriate than a general MIRT model. For instance, the model that the National Assessment of Educational Progress operationally employs uses the multiunidimensional model as one component. Moreover, overparameterization associated with the MIRT model makes the model identifiable only when \(\alpha_j\) is constrained (Béguin & Glas, 2001). The identification conditions are met for the case where every item loads on one dimension only and the association between the dimensions is modeled by a correlation matrix. This is exactly the multiunidimensional model discussed in this article.

Hierarchical Models for Bayesian Inference

To allow for all sources of uncertainty, it is useful to consider IRT models from a hierarchical perspective, in which each model parameter is given a prior distribution (e.g., Mislevy, 1986). Bayesian inference for such models is convenient because efficient computational algorithms (e.g., Markov chain Monte Carlo [MCMC]) can be used (Gelman, Carlin, Stern, & Rubin, 2003).

Hierarchical Multiunidimensional IRT Model

With the probabilistic model specified as in Equation 2, the likelihood function is

\[
f(y|\theta, \xi) = \prod_{v=1}^{m} \prod_{i=1}^{n} \prod_{j=1}^{k_v} p_{vij}^{y_{vij}} (1 - p_{vij})^{1-y_{vij}}.
\]
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 the correlation matrix $\Sigma$ can be easily transformed from $\Sigma^*$ using $r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}} (i \neq j)$. An inverse Wishart distribution is assumed for $\Sigma^*$ so that

$$p(\Sigma^*; I, 2) = \left(2^m \pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma \left(\frac{3-i}{2}\right)\right)^{-1} |\Sigma^*|^{m+3/2} \exp \left\{-\frac{1}{2} (tr(\Sigma^*-1))\right\}$$

(5)

where $I$ is an $m \times m$ identity matrix, and 2 is the degrees of freedom. Let $\xi_{vj} = (\alpha_{vj}, \gamma_{vj})$ denote the vector of item parameters for the $j$th item of the $v$th subtest. We consider flat priors for $\alpha_{vj}$ and $\gamma_{vj}$, that is, $p(\alpha_{vj}) \propto 1$ and introduce random variables $Z_{vij}$ so that $Z_{vij} \sim N(\alpha_{vj}\theta_{vj} - \gamma_{vj}, 1)$ and $y_{vij} = \begin{cases} 1, & \text{if } Z_{vij} > 0 \\ 0, & \text{if } Z_{vij} \leq 0 \end{cases}$. Thus, the joint posterior distribution of $(\theta, \xi, Z, \Sigma^*)$ is given in hierarchical form by

$$p(\theta, \xi, Z, \Sigma^* | y) \propto f(y | Z) p(Z | \theta, \xi) p(\xi) p(\theta | \Sigma) p(\Sigma^*)$$

(6)

Because it is difficult to obtain the normalizing constants for Equation 6, the Gibbs sampling procedure is adopted to simulate random samples of $\theta$ and $\xi$ from the respective posterior distributions (see Lee, 1995, for a detailed derivation and illustration of the procedure).

**Hierarchical Unidimensional IRT Model**

For the unidimensional IRT model, we follow Albert’s (1992) procedure. That is, let $\xi_j = (\alpha_j, \gamma_j)$ denote the vector of parameters for the $j$th item; let $\theta = (\theta_1, \ldots, \theta_n)$ denote the vector of ability parameters; and let $\xi = (\xi_1, \ldots, \xi_n)$, denote the vector of all item parameters. In this case, the likelihood function can be written as

$$f(y | \theta, \xi) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}},$$

(7)

where $p_{ij}$ is as defined in Equation 3. Assume $\theta_1, \ldots, \theta_n$ are independent and identically distributed samples from $N(0, 1)$, and assume the item discriminating parameter $\alpha_j$ is positive and $p(\gamma_j) \propto 1$. Thus, after introducing latent continuous random variables $Z_{ij}$ so that $Z_{ij} \sim N(\alpha_j \theta_j - \gamma_j, 1)$ and $y_{vij} = \begin{cases} 1, & \text{if } Z_{vij} > 0 \\ 0, & \text{if } Z_{vij} \leq 0 \end{cases}$, the posterior distribution of $(\theta, \xi, Z)$ is given by

$$p(\theta, \xi | y) \propto f(y | Z) p(Z | \theta, \xi) p(\theta) p(\xi).$$

(8)
As before, the Gibbs sampling procedure can be used to obtain random samples of parameters of interest from the posterior distribution; hence, inferences can be made accordingly.

**Bayesian Model Choice Techniques**

It is common to use Bayes factors in the Bayesian framework to compare models (Gelman et al., 2003). However, due to the use of noninformative priors in the hierarchical models presented above, Bayes factors do not yield meaningful results and hence are not considered. Rather, we adopt Bayesian deviance and posterior predictive model checks for comparing the hierarchical models.

**Bayesian Deviance**

The deviance information criterion (DIC) was introduced by Spiegelhalter, Best, and Carlin (1998), who generalized the classical information criteria to one that is based on the posterior distribution of the deviance. This criterion is defined as $\text{DIC} = \tilde{D} + p_D$, where $\tilde{D} = E_y[D] = E(−2 \log L_{\theta|y}(\theta))$ is the posterior expectation of the deviance (with $L_{\theta|y}$ being the likelihood function, i.e., $L_{\theta|y}(y|\theta) = \prod p(y|\theta)^y(1 − p(y|\theta))^{1−y}$, where $\theta$ denotes all model parameters, and $p(\theta)$ is as defined in Equation 2 for the 2P multiunidimensional model) and $p_D = E_y[D] − D(\theta)$ is the effective number of parameters (Carlin & Louis, 2000). Furthermore, $D(\theta) = −2 \log (L_{\theta|y}(y|\theta))$, where $\theta$ is the posterior mean. To compute 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, and $\tilde{D}$ can be approximated as $\tilde{D} = \frac{1}{G} \sum_{g=1}^{G} L_{\theta|y}(y|\theta^{(g)}, \xi^{(g)})$. Generally, the more complicated models tend to provide better fits. 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.

**Posterior Predictive Model Checks**

Among the methods proposed for model checking, posterior predictive checking is easy to carry out and interpret in spite of being conservative because the model is assessed by checking the observed data against data generated using the posterior estimates of the parameters under the model (Sinharay & Stern, 2003). The basic idea is to draw simulated values from the posterior predictive distribution of replicated data, $y^{\text{rep}}$, $p(y^{\text{rep}}|y) = \int \int p(y^{\text{rep}}|\xi, \theta)p(\xi, \theta|y)d\xi d\theta$, and compare the simulated values to the observed data $y$. If the model fits, then replicated data generated under the model should look similar to the observed data. A test statistic $T(y, (\xi, \theta))$...
is chosen to define the discrepancy between the model and the data. If there are \( Q \) simulations from the posterior distribution of \((\xi, \theta)\), one \( y^{\text{rep}} \) can be drawn from the predictive distribution for each simulated \((\xi, \theta)\), so there are \( Q \) draws from the joint posterior distribution \( p(y^{\text{rep}}, \xi, \theta|y) \). It is then easy to compare the realization-based test statistics \( T(y, (\xi^q, \theta^q)) \) with the predictive test statistics \( T(y^{\text{rep}q}, (\xi^q, \theta^q)) \) by plotting the pairs on a scatter plot. Alternatively, one can calculate the probability or posterior predictive \( p \) value (PPP value; Sinharay & Johnson, 2003) that the replicated data could be more extreme than the observed data: \( p_B = \Pr(T(y^{\text{rep}}, (\xi, \theta)) \geq T(y, (\xi, \theta)|y) ). \)

**Simulation Studies**

To assess the performance of the multiunidimensional model relative to the unidimensional model, simulation studies were conducted to represent tests with different dimension structures. In particular, we considered four situations described below. A total of 82 item parameters, \( a \) and \( \gamma \), were randomly generated from uniform distributions so that \( p(a) > 0 \), and \( p(\gamma) \propto 1 \), and they were used in all the following simulations so that 1,000 responses to these items were simulated.

In Simulation 1, all items were assumed to measure the same latent ability so that the ability \( y \) is from a univariate normal distribution, \( y_i \sim N(0,1) \). The responses were simulated based on the probability as defined in Equation 3. In Simulation 2, the test of interest consisted of two subtests (each having 16 and 25 items) with the latent abilities \( y_1 \) and \( y_2 \) assumed to be highly correlated with a correlation of .8. Hence, \( \theta_i = (y_{1i}, y_{2i}) \) was generated from a multivariate normal distribution, \( \theta_i \sim N_2(0, \Sigma) \), where \( \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \). The responses were then simulated from the probabilities as defined in Equation 2, where \( m = 2 \). In Simulation 3, we assumed a lower correlation between the two abilities (.6), and thus \( \Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \). Otherwise, this simulation was identical to Simulation 2. Last, Simulation 4 considered the situation where the abilities were uncorrelated. In other words, an examinee’s ability to answer items in one cluster was assumed to be independent of his or her ability to answer items in the other. In this case, we generated \( \theta_i = (\theta_{1i}, \theta_{2i}) \) from a multivariate normal distribution, \( \theta_i \sim N_2(0, I) \), where \( I \) is the identity matrix. Although this scenario would be unusual in practice, it is illustrative because it represents the extreme situation where the unidimensional model does not perform well.

For each of the four scenarios, we considered three model implementations. First, we fit the unidimensional IRT model together (for all items). Second, we fit the unidimensional model separately on each of the two subtests. Finally, we fit the multiunidimensional model. Each implementation involved 10 replications. Gibbs
sampling was implemented, where 7,000 iterations were obtained, with the first 2,000 as burn-in. Convergence was assessed using R statistics (Gelman et al., 2003), and the Brooks-Gelman multivariate potential scale reduction factor (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 (EAPs) of the Gibbs samples. To examine the item parameter recovery in each case, root-mean-squared differences between true and estimated item parameters were obtained in each replication, and their averages were used to describe how accurately the parameters were estimated for the four simulations. For all 41 item discrimination parameters, $\alpha$, we used $\alpha_1$ to denote those for the first 16 items and $\alpha_2$ for the remaining 25 items. Separate implementation of the unidimensional model produced two sets of item EAPs and therefore was not included in the item parameter comparison.

As shown in Table 1, in the first three simulations, the two models did not differ much in their average root-mean-squared differences between the actual and recovered item parameters. However, in Simulation 4, the multiunidimensional model had smaller average root-mean-squared difference values for $\alpha_1$ and $\gamma$, and was obviously better than the unidimensional model. It has to be noted that for the unidimensional model, the root-mean-squared difference for $\alpha_1$ increased markedly from Simulation 1 through Simulation 4. In other words, $\alpha_1$ is less likely to get recovered well as the actual situation departs more from unidimensionality. As to the recovery of the person ability parameters, we examined correlations between true and estimated values because in the educational testing situation, one is often interested only in the relative values of $\theta$ instead of the true values for different examinees. The correlations for the three approaches in the four simulations were averaged for 10 replications and summarized in Table 2. Comparatively, the multiunidimensional model recovered the abilities as well as the unidimensional model in Simulation 1 and better than the unidimensional model in Simulations 2, 3, and 4. The unidimensional model performed the worst in Simulation 4, especially in recovering $\theta_1$. This is in accordance with the observation that $\alpha_1$ was recovered less well than the other item parameters. Separate implementation of the unidimensional model was no better than the multiunidimensional model in ability parameter recovery in any of the four simulated scenarios, even in Simulation 4 where the abilities were not related at all.

Next, we compared the multiunidimensional model with the separate implementation of the unidimensional model regarding the recovery of the dimensional structure. The correlations between the ability EAPs, $\hat{\theta}_1$ and $\hat{\theta}_2$, as well as the EAPs of the correlation parameter, $\hat{\rho}$, in the multiunidimensional model were obtained in each of the 10 replications, and their averages are shown in Table 3. These values were compared with the true correlations in each simulation to assess how well the structure of the ability dimensions was recovered. It can be easily seen from the table that both $\hat{\rho}$ and the correlations between $\hat{\theta}_1$ and $\hat{\theta}_2$ for the
multiunidimensional model were close to the true correlation values in all four simulations. On the other hand, separate implementation of the unidimensional model worked well only in Simulation 4 when there was no correlation between the true latent abilities. In the other situations, the estimated correlations showed smaller values than the true correlations. A close examination of the correlations between the ability EAPs for separate implementation of the unidimensional model in the four simulations (i.e., .6964, .5647, .4076, and -.0036 from Table 3) indicates that although the correlation between the posterior expectations of $\theta_1$ and $\theta_2$ decreases as

**Table 1**

Average Root-Mean-Squared Differences Between the Actual and Estimated Item Parameters for the Unidimensional and Multiunidimensional Models in the Four Simulations (10 Replications)

<table>
<thead>
<tr>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
<th>Simulation 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unidimensional model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_1$</td>
<td>.0583</td>
<td>.0755</td>
<td>.1459</td>
</tr>
<tr>
<td>$a_2$</td>
<td>.0616</td>
<td>.0557</td>
<td>.0592</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>.0539</td>
<td>.0539</td>
<td>.0574</td>
</tr>
<tr>
<td><strong>Multiunidimensional model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_1$</td>
<td>.0624</td>
<td>.0693</td>
<td>.0678</td>
</tr>
<tr>
<td>$a_2$</td>
<td>.0616</td>
<td>.0608</td>
<td>.0608</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>.0548</td>
<td>.0557</td>
<td>.0529</td>
</tr>
</tbody>
</table>

**Table 2**

Correlations Between the Actual and Estimated Person Abilities for the Unidimensional Model, Separate Implementation of the Unidimensional Model, and the Multiunidimensional Model in the Four Simulations (10 Replications)

<table>
<thead>
<tr>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
<th>Simulation 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unidimensional model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\theta}_1$</td>
<td>.9064</td>
<td>.8189</td>
<td>.8807</td>
</tr>
<tr>
<td>$\hat{\theta}_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Separate implementation of the unidimensional model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\theta}_1$</td>
<td>.7924</td>
<td>.7915</td>
<td>.7903</td>
</tr>
<tr>
<td>$\hat{\theta}_2$</td>
<td>.8664</td>
<td>.8679</td>
<td>.8700</td>
</tr>
<tr>
<td><strong>Multiunidimensional model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\theta}_1$</td>
<td>.9076</td>
<td>.8551</td>
<td>.8289</td>
</tr>
<tr>
<td>$\hat{\theta}_2$</td>
<td>.9090</td>
<td>.8883</td>
<td>.8807</td>
</tr>
</tbody>
</table>
the true correlation decreases, the magnitude of the decrease is not proportional to the decrease in the true correlation. In general, based on the pattern, the higher the correlation between the true \( \theta \)'s, the greater the estimated value shrinks toward zero.

In summary, as far as parameter recovery is concerned, both the unidimensional model and the multiunidimensional model perform well when the latent abilities are perfectly correlated. When the abilities are highly correlated, the unidimensional model, when compared with the multiunidimensional model, works equally well in recovering item parameters but less well in recovering person abilities. And when the abilities are not highly correlated, the multiunidimensional model is obviously more accurate in recovering all model parameters. Moreover, it accurately recovers the dimensional structure in all four simulated situations as well. Overall, the results from the simulation studies confirm that the unidimensional model can be used only when the ability dimensions are highly correlated. The multiunidimensional model, on the other hand, is applicable in situations with various dimensional structures (including unidimensionality), and hence it is considered to be more flexible. Separate implementation of the unidimensional model does not prove to be better than either of the former two models in the simulated scenarios.

**An Example With CBASE English Data**

In real test situations, the true latent structure is not necessarily known. Hence, model comparison is necessary to determine which model provides a better representation of the data. We use CBASE English data as an illustration, which consist of 1,231 randomly selected examinees. The English test is further organized into levels of increasing specificity by two clusters, that is, writing and reading/literature. If one conducts a confirmatory analysis, the nature of the test limits the candidate models to be either a unidimensional IRT model or a multiunidimensional model.
If the unidimensional model is not adequate for the overall data, that is, the two clusters are measuring distinct ability dimensions, a unified general ability measure cannot describe the specific true latent traits necessary for the writing and reading/literature clusters.

Because the multiunidimensional model can accurately recover the structure of the ability dimensions, it was implemented via the Gibbs sampling procedure, and the correlation, $\rho$, between the estimated abilities for the two clusters is estimated to be .8175 with a variance of .0315. The posterior density for this parameter suggests that an approximate 95% credible interval for the true correlation is (.76, .88). This indicates a high correlation between the two specific abilities. One may argue that the two clusters measuring highly correlated ability dimensions actually supports the use of the unidimension model. Therefore, a unidimensional model may be better than the multiunidimensional model for the CBASE data. Model comparisons are thus carried out to check this presumption.

To carry out model comparisons using Bayesian model checking procedures for the CBASE English data, we considered three candidate models, two of which are as defined in the IRT Models section with Model 1 for single ability and Model 2 for two abilities with $\theta_i \sim N(0, \Sigma)$. We added a third model assuming the prior for the two abilities as $\theta_i \sim N(0, I)$, which is similar to implementing the unidimensional model separately for the two clusters, as described earlier.

The Bayesian deviance results are summarized in Table 4, where Model 1 is shown to have a smaller DIC value compared to Model 3. However, Model 1 has the largest posterior expectation of the deviance among the three models. The posterior densities of the deviance obtained from these models are plotted in Figure 1. Models 2 and 3 are quite similar, whereas Model 1 (the dotted line) has uniformly larger deviance than the other two. The different results with posterior deviance and DIC for Model 1 are due to the fact that DIC penalizes for a large effective number of parameters, $p_D$. As shown in the last column of Table 4, Model 1 has the fewest effective number of parameters. Generally, after controlling for an effective number of parameters, Model 2 seems to be the best choice among the three candidate models, although Model 1 is not too bad compared with Model 3. Thus, based on the Bayesian DIC criterion, there is evidence in favor of the more general

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>$\bar{D}$</th>
<th>$D(\bar{\theta})$</th>
<th>$p_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>56393</td>
<td>55309</td>
<td>54225</td>
<td>1083.60</td>
</tr>
<tr>
<td>Model 2</td>
<td>56270</td>
<td>54823</td>
<td>53376</td>
<td>1447.10</td>
</tr>
<tr>
<td>Model 3</td>
<td>56632</td>
<td>54840</td>
<td>53049</td>
<td>1791.30</td>
</tr>
</tbody>
</table>

Note: DIC = deviance information criterion.
multiunidimensional model, which in turn suggests less adequacy of the unidimensional model in this context.

Next, we implemented the posterior predictive model checking procedure for comparing the three candidate models. To do so, a test statistic must be chosen for describing the discrepancy between the model and data. For this analysis, we adopted two discrepancy measures: the proportion correct for the items and the odds ratio for measuring association among item pairs. The former is considered to be less powerful but can be useful for checking the computations, whereas the latter, $T(y) = OR_{ij} = \frac{n_{11}n_{00}}{n_{10}n_{01}}$, is powerful for detecting unidimensionality in the data (SinhaRay & Johnson, 2003). For each fitted model, based on each pair of $\xi, \theta$.
Figure 2
Boxplots for Item Proportion Correct With the Three Candidate Models
samples, a $y^{rep}$ was simulated and the replicated discrepancy $T(y^{rep})$ was computed to be compared with the actual discrepancy. The tail-area PPP values ($p_B$) are estimated as the proportion of the simulated samples for which $T(y^{rep}) \geq T(y)$, that is, $p_B = \sum_{i=1}^{L} I(T(y^{rep})) \geq T(y)$.

Figure 2 shows boxplots of the replicated proportion correct for each item with the three fitted models. In the figure, the dots denote the observed proportion correct for each item. It can be seen that with all three models, the observed proportion-correct values are close to the median of the replicated values. A close examination of all the tail-area PPP values suggest that the replicated data based on each of the three candidate models are not extreme when compared with the observed data. Thus, when using item proportion correct as the discrepancy statistic, the three models fit similarly well.

In contrast to the proportion-correct results, there are noticeable differences with the odds ratios. Similar to the plots in Sinharay and Johnson (2003) and Sinharay (2005), Figure 3 summarizes the extreme PPP values for the odds ratios with each model. Here we used $\alpha = .05$ as the critical level, so the PPP value larger than
.975 is denoted using a plus sign and the PPP value smaller than .025 is denoted using a cross sign. Because odds ratios are based on the responses to any pair of items, each plot is symmetrical about its diagonal; that is, the upper diagonal carries the same information as the lower diagonal. Hence, the upper diagonal is left blank for simplicity.

From the figure, it is immediately clear that with much fewer extreme replicated odds ratios, Models 1 and 2 perform much better than Model 3. Model 1 has the same number of extreme PPP values as Model 2. However, careful examination of the plot reveals quite different prediction errors. We know that there are 41 English items, with items 1-16 in the writing cluster and items 17-41 in the reading/literature cluster. Model 1 mostly underpredicts the odds ratios between items within the two clusters and overpredicts the odds ratios between the items from each cluster. The underpredictions are outcomes of associations among the items beyond the model’s prediction capacity, and the overpredictions are because some of the examinees show higher ability for one dimension and lower ability for the other, which is not in accordance with the assumption of the unidimensional model that all examinees at a specific ability level should get all items correct within their ability level. On the other hand, Model 2 does not show as many underpredictions within the two clusters, especially in writing. It performs similarly for items in the reading/literature cluster as Model 1. However, Model 2 underpredicts a few more item pairs between the two clusters. Note that these underpredictions are simply because the examinee’s two abilities are more similar than Model 2 specifies; that is, high ability in one dimension does not have to be associated with similarly high ability in the other.

Overall, with the posterior predictive model checking method, the multiunidimensional model describes the data as well as the unidimensional model. When using proportion correct as the discrepancy statistic, all models show similar goodness of fit. Furthermore, in agreement with Sinharay & Johnson’s (2003) conclusion, odds ratios are shown to be more informative than the item proportion correct as far as these three models are concerned. This measure suggests Models 1 and 2 are better, but with different prediction errors.

In practice, one might want to know the practical consequences of model misfit when using the simpler unidimensional IRT model (Model 1) instead of the complicated multiunidimensional model (Model 2). One of the primary purposes in many test situations lies in score reporting. Hence, if Model 1 results in a substantial difference (from Model 2) in the person ability estimates, then the misfit of the former model is of substantial practical concern. Figure 4 compares the posterior means and standard deviations (SDs) for each examinee’s abilities using Model 1 and Model 2 for the CBASE English data. It is clear that the unidimensional model consistently underestimates the posterior SD. The posterior means differ as well. The unidimensional model is more likely to overestimate the value for above-average scores and underestimate that for below-average scores. Moreover, the
general ability ($\theta$) does not represent the writing ability ($\theta_1$) well for all 1,231 examinees, although it is fairly close to the reading ability ($\theta_2$). Consequently, with these observations, and the fact that the two models do not differ much in computing time (using MATLAB programs), some general guidelines can be drawn: If we think that these magnitudes of differences between the two models are significant, we should use the more complicated multiunidimensional model. Otherwise, we can go with the simpler unidimensional model.

**Conclusions**

For tests consisting of multiple subtests, the unidimensional model is believed to work well when the subtests are known to measure the same underlying ability.
It is not appropriate for tests involving distinct ability dimensions. The multiunidimensional model, on the other hand, is relatively robust to the true process being unidimensional.

From the simulation studies comparing a multiunidimensional model with a unidimensional model for tests with different dimensionality structures, we show that comparatively, the multiunidimensional model is more flexible and works well even when the actual test is unidimensional, whereas the unidimensional model works well only when all test items are strictly measuring the same ability. Moreover, the multiunidimensional model accounts for the relationship between the abilities and thus produces more reliable subscores than separate implementation of a unidimensional model.

In most instances, due to the lack of a satisfactory index for assessing the dimensionality assumption, the structure of the ability dimensions is not clear. Under such circumstances, we should use the more general multiunidimensional IRT model unless there is compelling evidence to suggest otherwise. Model comparisons for a subset of CBASE English data further illustrate this point. Both Bayesian DIC and the posterior predictive model checks indicate that the multiunidimensional IRT model provides a goodness of fit similar to, if not better than, the unidimensional model. It must be noted that in the CBASE data, the two subtests are found to be highly correlated. This further suggests that even with test items measuring similar abilities, the unidimensional model does not provide a better description of the data than the multiunidimensional model. Therefore, the multiunidimensional model offers a better way to represent test situations not realized in unidimensional models.

In this article, we discussed the practical consequences of model misfit but did not delve into that; further studies using more rigorous analyses (Hambleton & Han, 2004) can be performed. Regarding predictive model checks, we adopted item proportion correct and odds ratios as two discrepancy measures. Other test statistics could also be considered, such as item test-item biserial correlations and observed score distribution. The choice of discrepancy measures is crucial with this method. Some measures may fail to detect the differences between models, such as item proportion correct as illustrated in the CBASE example. In addition, 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 & Chib (1993). Additionally, Bayes factors, known as the counterpart of likelihood ratio tests, provide more reliable and powerful model comparisons in the Bayesian framework and thus can be considered if one is using informative priors. The current study focused on 2P normal ogive IRT models with no more than two ability dimensions. Future studies can consider other IRT models, such as 3P models, logistic models, or models with 3 or more abilities.
References


