A Multidimensional Partial Credit Model With Associated Item and Test Statistics: An Application to Mixed-Format Tests

Lihua Yao and Richard D. Schwarz, CTB/McGraw-Hill

Multidimensional item response theory (IRT) models have been proposed for better understanding the dimensional structure of data or to define diagnostic profiles of student learning. A compensatory multidimensional two-parameter partial credit model (M-2PPC) for constructed-response items is presented that is a generalization of those proposed to date along with a compensatory multidimensional three-parameter logistic model for multiple-choice data (M-3PL). Estimation of these models using Markov chain Monte Carlo methods is discussed. To further evaluate these models and characterize item and test functioning, multidimensional representations of statistics such as information, difficulty, and discrimination for the M-3PL and M-2PPC are presented. Many assessment programs have a mixture of item types in which multiple choice and constructed response are administered together. An example is presented in which the dimensional structure of a test containing mixed item types is examined. Goodness-of-fit testing under various model formulations and derived statistics are discussed. *Index terms: item response theory, partial credit model, MIRT, item information, item statistics, MCMC*

Introduction

Many types of assessments currently contain a mixture of multiple-choice and constructed-response tasks that require different item response models. A survey from Lane (2005) stated that 63% of the state assessment uses both multiple-choice items and constructed-response items. Constructed-response items are hypothesized to measure concepts and skills at greater depth than multiple-choice items (Ercikan et al., 1998). These items can be costly but can reduce the test length to achieve the same goal, even in computer adaptive testing (De Ayala, 1989, 1992). A common hypothesis is that constructed-response items measure traits that are different from multiple-choice items (Traub, 1993). An item designed to measure one trait may also require some level of other traits. Multidimensional item response theory (MIRT) is a generalization of unidimensional IRT that describes the interaction between a person and a task where the characteristics of the person are described by a vector of constructs. The structure of mixed-format tests is examined using the multidimensional three-parameter logistic (M-3PL) model applied to multiple-choice items and the multidimensional two-parameter partial credit model (M-2PPC) for constructed-response tasks.

For the unidimensional case, IRT models have been proposed for constructed-response items, such as the partial credit or graded response model, and the Rasch or three-parameter logistic models

Applied Psychological Measurement, Vol. 30 No. 0, Month 2006, 1–25
DOI: 10.1177/0146621605284537

in the case of multiple-choice items. Many different parameterizations of multidimensional models have been formulated and applied for a variety of purposes, such as examining the underlying structure of data (Ackerman, Gierl, & Walker, 2003), checking model assumptions, or defining diagnostic profiles of student learning (Walker, & Beretvas, 2003). Multidimensional versions of the normal-ogive, Rasch, three-parameter models have also been proposed (Ackerman, 1994; Adams, Wilson, & Wang, 1997; Kelderman, 1996; Reckase, 1985) for dichotomous test data.

Several software programs (e.g., TESTFACT [Wilson, Wood, & Gibbons, 1991; NOHARM [Fraser & McDonald, 1988]) implement these MIRT models for dichotomous data. For polytomous data, POLYFACT (Muraki, 1999) implements the multidimensional graded response model or the generalized partial credit model. It uses a marginal maximum likelihood estimation strategy and imposes a confirmatory parameter structure in which the score levels of constructed-response items are associated with a particular dimension, and it does not calibrate M-3PL models. The MicroFACT program (Waller, 2002) uses exploratory factor analysis and requires users to input the guessing parameters. BMIRT (Yao, 2003), which uses Markov chain Monte Carlo (MCMC) methods, does not impose these sorts of constraints. It implements the M-2PPC and can be used concurrently with the M-3PL model for mixed formats.

MIRT models can be characterized as either being compensatory or noncompensatory (Ackerman, 1989). The compensatory version allows the dimensions to interact, with high ability on one dimension, which can potentially compensate for lower ability on a second dimension. For the noncompensatory version, one must be proficient in both abilities to obtain a higher score. In this article, focus is given to the compensatory multidimensional version of M-3PL and M-2PPC models.

The overall fit of a multidimensional model and comparison with others of interest can be tested using customary fit statistics such as Akaike's (1987) information criterion (AIC) and difference chi-square. The AIC statistic imposes a penalty for adding additional parameters to the model. The difference chi-square can be used to compare nested models to select a preferred candidate. The dimensional structure of tests that contain mixed item types was evaluated for the M-3PL and M-2PPC models using these goodness-of-fit statistics.

In addition to overall measures of model fit, multidimensional statistics have been used within an IRT framework to further describe the characteristics of items and tests. Multidimensional representations of statistics for dichotomous data such as information, difficulty, discrimination, and their graphical representations have been developed for the multidimensional two-parameter logistic (M-2PL) and M-3PL models (Ackerman, 1994, 1996; Reckase, 1985; Reckase & McKinley, 1991). To further evaluate these models and characterize item and test functioning, multidimensional representations of statistics for information, difficulty, and discrimination are developed here for the M-2PPC model along with their interpretation. Because assessments containing mixed formats are examined here, multidimensional statistics for multiple-choice items are also presented. For the purpose of comparison among different models, unidimensional and classical statistics are provided together with their multidimensional counterparts. For instance, item discrimination for constructed-response items can be compared for both unidimensional and multidimensional models.

Multidimensional Models and Their Estimation

Suppose there are N examinees and J test items. The observable item response data are contained in a matrix $X = \{X_{ij}\}$, where $i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, J$. For convenience, let the responses of the ith examinee be a vector $\vec{X}_i = (X_{i1}, , X_{iJ})$. The ability parameters for examinees are

$$\theta = (\vec{\theta}_1, \ldots, \vec{\theta}_N)^T, \tag{1}$$

such that each $\vec{\theta}_i$ for $i = 1, 2, \ldots, N$ is a vector of dimension D, where D is the number of subscales or the number of dimensions hypothesized.

For a dichotomous item j, the probability of a correct response to item j for an examinee with ability $\vec{\theta}_i$ for the M-3PL model is

$$P_{ij1} = P(x_{ij} = 1 \mid \vec{\theta}_i, \vec{\beta}_j) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{(-\vec{\beta}_{2j} \odot \vec{\theta}_i^T + \beta_{1j})}}, \tag{2}$$

where

- $x_{ij} = 0$ or 1 is the response of examinee i to item j.
- $\vec{\beta}_{2j} = (\beta_{2j1}, \ldots, \beta_{2jD})$ is a vector of dimension D of item discrimination parameters.
- β_{1j} is the scale difficulty parameter.
- β_{3j} is the scale guessing parameter.
- $\vec{\beta}_{2j} \odot \vec{\theta}_i^T$ is a dot product of two vectors.[1]

The parameters for the jth item are

$$\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j}). \tag{3}$$

For a polytomous item j, the probability of a response $k - 1$ to item j for an examinee with ability $\vec{\theta}_i$ is given by the multidimensional version of the partial credit model (M-2PPC):

$$P_{ijk} = P(x_{ij} = k - 1 \mid \vec{\theta}_i, \vec{\beta}_j) = \frac{e^{(k-1)\vec{\beta}_{2j} \odot \vec{\theta}_i - \sum_{t=1}^{k} \beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{((m-1)\vec{\beta}_{2j} \odot \vec{\theta}_i^T - \sum_{t=1}^{m} \beta_{\delta_t j})}}, \tag{4}$$

where

- $x_{ij} = 0, \ldots, K_j - 1$ is the response of examinee i to item j.
- $\vec{\beta}_{2j} = (\beta_{2j1}, \ldots, \beta_{2jD})$ is a vector of dimension D of item discrimination parameters.
- $\beta_{\delta_k j}$ for $k = 1, 2, \ldots, K_j$ are the threshold or alpha parameters, $\beta_{\delta_1 j} = 0$, and K_j is the number of response categories for the jth item.

The parameters for the jth item are

$$\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{\delta_2 j}, \ldots, \beta_{\delta_{K_j} j}). \tag{5}$$

Let

$$P_{ij} = P(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j) = P_{ij1}^{1_{(X_{ij}=1)}} (1 - P_{ij1})^{1_{(X_{ij}=0)}}, \tag{6}$$

for an M-3PL item or

$$P_{ij} = P(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j) = \prod_{k=1}^{K_j} P_{ijk}^{1_{(X_{ij}=k-1)}}, \tag{7}$$

for an M-2PPC item, where

$$1_{(X_{ij}=k)} = \begin{cases} 1 & \text{if } X_{ij} = k \\ 0 & \text{otherwise} \end{cases}.$$

The item parameters for all test items are

$$\beta = (\vec{\beta}_1, \ldots, \vec{\beta}_j, \ldots, \vec{\beta}_J)^T. \tag{8}$$

The likelihood equation is

$$P(X \mid \theta, \beta) = \prod_{i=1}^{N} P(\vec{X}_i \mid \vec{\theta}_i, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j). \tag{9}$$

Let $P(\theta \mid \lambda)$ be the probability distribution for an examinee population with ability θ, given parameter λ. For example, if θ is assumed to be normally distributed, then $\lambda = (\mu, \sigma)$, where μ is the mean and σ is the variance and covariance matrix. The joint posterior distribution can be written as

$$P(\theta, \beta, \lambda \mid X \propto P(X \mid \theta, \beta, \lambda) P(\theta \mid \beta, \lambda) P(\beta) P(\lambda) \tag{10}$$

$$= P(X \mid \theta, \beta) P(\theta \mid \lambda) P(\beta) P(\lambda), \tag{11}$$

where

$$P(X \mid \theta, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(X_{ij} \mid \vec{\theta}_i, \vec{\beta}_j), \tag{12}$$

$$P(\theta \mid \lambda) = \prod_{i=1}^{N} P(\vec{\theta}_i \mid \lambda), \tag{13}$$

with the assumption that the examinee's response to an item depends on the examinee's ability and the item parameters.

Like the one-dimensional IRT models, MIRT models are underidentified. The probability of a response will not change after certain linear transformations on parameters. The metric or scale is imposed in the estimating software.

Estimation can be problematic with multidimensional models due to the inclusion of many additional parameters compared with the unidimensional case. MCMC methods sample from the posterior distribution, which avoids taking the derivative that can be problematic when a high number of dimensions are present. The parameters (θ, β, λ) are estimated using the Metropolis-Hastings algorithm that samples from the joint posterior $P(\theta, \beta, \lambda \mid X)$. Metropolis-Hastings is a general term for Markov chain simulation methods, which are useful for drawing samples from appropriate distributions and then correcting those draws to better target the posterior distribution (Patz & Junker 1999a, 1999b). A Bayesian formulation of multivariate item response theory was implemented in a computer program called BMIRT (Yao, 2003, 2004a). This program uses MCMC methods to estimate the item, ability, and population parameters for multidimensional, multigroup models for dichotomous and polytomous data, which can be used in either exploratory or confirmatory modes. The parameter recovery studies of BMIRT can be found in Patz and Yao (2003) and Yao and Boughton (2005a, 2005b).

Multidimensional Items and Test Statistics

This section presents formulations of item and test statistics for the M-2PPC model that are conceptually similar to those that have been developed for their dichotomous (i.e., multiple-choice) counterparts. For completeness and comparative purposes, these multidimensional formulations for item difficulty have also been applied to the unidimensional case.

Multidimensional Item and Test Information

Within IRT, information functions are used to determine the measurement precision for a designated level of ability. In the multidimensional case, information corresponds to the composite of abilities in space. The direction in the space must be considered when evaluating the information provided by the item. Reckase and McKinley (1991) presented a method for computing information for the M-2PL model. Ackerman (1994) extended the version that Green (1990) developed, which accounts for the lack of local independence among factors when the information is estimated for a particular direction using the M-2PL model. Multidimensional item and test information using these two definitions will be derived for the M-2PPC model in the next section.

First Definition of Item Information

Multidimensional item and test information for the M-2PL, as presented in Reckase and McKinley (1991), was applied to the M-3PL and extended to the M-2PPC model. This definition has the advantage of providing an approximation to information for a single item and for test information. For convenience, index i, which represents examinee i, is dropped in subsequent notation. For example, the probability of answering item j in equation (7) can be defined as

$$P_j = P_j(\vec{\theta}) = P(x_j \mid \vec{\theta}, \vec{\beta}_j) = \prod_{k=1}^{K_j} P_{jk}^{1(x_j = k-1)}. \tag{14}$$

Here, $P_{jk} = P_{jk}(\vec{\theta})$is defined by equation (4) for M-2PPC items.

The item information function is defined by

$$I_j(\vec{\theta})(\vec{\alpha}) = \sum_{k=1}^{K_j} \frac{(\nabla_\alpha P_{jk}(\vec{\theta}))^2}{P_{jk}}, \tag{15}$$

where $\nabla_\alpha P_{jk}(\vec{\theta})$ is the directional derivative of function $P_{jk}(\vec{\theta})$ in the direction $\vec{\alpha}$ and can be expressed as follows:

$$\nabla_\alpha P_{jk}(\vec{\theta}) = \frac{\partial \vec{P}_{jk}}{\partial \theta} \odot (\vec{\cos}\, \alpha)^T. \tag{16}$$

Here, $\vec{\theta} = (\theta_1, \ldots, \theta_D)$, and $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)$.

- For each M-3PL item j, note that P_{j1} from equation (2) can also be expressed as follows:

$$P_{j1} = \frac{1 + \beta_{3j} e^{(-\vec{\beta}_{2j} \odot \vec{\theta}^T + \beta_{1j})}}{1 + e^{(-\vec{\beta}_{2j} \odot \vec{\theta}^T + \beta_{1j})}}, \tag{17}$$

and

$$\frac{\partial \vec{P}_{j1}}{\partial \theta} = \frac{(1 - P_{j1})}{1 + e^{(-\vec{\beta}_{2j} \odot \vec{\theta}^T + \beta_{1j})}} \vec{\beta}_{2j}. \tag{18}$$

Therefore, after some computation,

$$I_j(\vec{\theta})(\vec{\alpha}) = \frac{(\nabla_\alpha P_j 1(\vec{\theta}))^2}{P_j 1(1 - P_j 1)} \tag{19}$$

$$= \frac{1}{P_j 1(1 - P_j 1)} \left(\frac{(1 - P_j 1)}{1 + e^{(-\vec{\beta}_2 j \odot \vec{\theta}^T + \beta_1 j)}} \right)^2 (\vec{\beta}_2 j \odot (\vec{\cos}\, \alpha)^T)^2 \tag{20}$$

$$= \frac{P_j 1(1 - P_j 1)}{(1 + \beta_3 j e^{(-\vec{\beta}_2 j \odot \vec{\theta}^T + \beta_1 j))^2}} (\vec{\beta}_2 j \odot \vec{\cos}\, \alpha)^T)^2. \tag{21}$$

- For each M-2PPC item j, note that the expectation is

$$E_j = \sum_{k=1}^{K_j}(k-1)P_{jk} = \frac{\sum_{k=1}^{K_j}(k-1)e^{(k-1)\vec{\beta}_{2j}\odot\vec{\theta}T - \sum_{t=1}^{k}\beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{((m-1)\vec{\beta}_{2j}\odot\vec{\theta}T - \sum_{t=1}^{m}\beta_{\delta_t j})}}. \tag{22}$$

Taking the log on both sides of equation (4) results in

$$\log P_{jk} = (k-1)\vec{\beta}_{2j}\odot\vec{\theta}\,T - \sum_{t=1}^{k}\beta_{\delta_t j} - \log\left(\sum_{m=1}^{K_j} e^{((m-1)\vec{\beta}_{2j}\odot\vec{\theta}T - \sum_{t=1}^{m}\beta_{\delta_t j})}\right). \tag{23}$$

Taking the derivative on both sides of equation (23) yields the following:

$$\frac{\partial\vec{P}_{jk}}{\partial\theta} = P_{jk}((k-1)-E_j)\vec{\beta}_{2j}. \tag{24}$$

Therefore,

$$I_j(\vec{\theta})(\vec{\alpha}) = \sum_{k=1}^{K_j} P_jk((k-1)-E_j)^2(\vec{\beta}_2 j\odot(\vec{\cos}\,\alpha)^T)^2 \tag{25}$$

$$= \left(\sum_{k=1}^{K_j} P_jk(k-1)^2 - E_j^2\right)(\vec{\beta}_2 j\odot(\vec{\cos}\,\alpha)^T)^2 I_j(\vec{\theta})(\vec{\alpha}) \tag{26}$$

$$= \sigma_j^2(\vec{\beta}_2 j\odot(\vec{\cos}\,\alpha)^T)^2, \tag{27}$$

where

$$\sigma_j^2 = \sum_{k=1}^{K_j}(k-1)^2 P_{jk} - E_j^2. \tag{28}$$

A Second Definition for Test Information

A second definition of information, as shown in Ackerman (1994) for M-2PL model, is developed here for the M-3PL and M-2PPC models. The alternative information measure takes the covariance among factors or traits into account. Note that with this formulation of test information, direction for the item is not used.

$$I_j(\vec{\theta}) = -E\frac{\partial^2\,\vec{\log}P_j}{\partial\theta^2} = -E\frac{\partial\frac{\partial\,\vec{\vec{\log}}P_j}{\partial\theta}}{\partial\theta}, \tag{29}$$

that is,

$$I_j(\vec{\theta}) = -E\begin{pmatrix} \frac{\partial^2\log P_j}{\partial\theta_1\partial\theta_1} & \frac{\partial^2\log P_j}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2\log P_j}{\partial\theta_1\partial\theta_D} \\ \frac{\partial^2\log P_j}{\partial\theta_1\partial\theta_2} & \frac{\partial^2\log P_j}{\partial\theta_2\partial\theta_2} & \cdots & \frac{\partial^2\log P_j}{\partial\theta_2\partial\theta_D} \\ \vdots & \vdots & \ddots & \\ \frac{\partial^2\log P_j}{\partial\theta_1\partial\theta_D} & \frac{\partial^2\log P_j}{\partial\theta_2\theta_D} & \cdots & \frac{\partial^2\log P_j}{\partial\theta_D\partial\theta_D}, \end{pmatrix}_{D\times D}, \tag{30}$$

where P_j is defined in equations (6) and (7). Taking log on both sides of equations (6) and (7),

$$\log P_j = 1_{(X_j=1)}\log P_{j1} + 1_{(X_j=0)}\log(1-P_{j1}) \tag{31}$$

for an M-3PL item and

$$\log P_j = \sum_{k=1}^{K_j} 1_{(X_j=k-1)} \log P_{jk} \tag{32}$$

for an M-2PPC item. Taking the derivative on both sides of equations (31) and (32), the following is obtained:

- For M-3PL item j,

$$\frac{\partial\, \vec{\log} P_j}{\partial\theta} = \left(\frac{1_{(X_j=1)}}{P_{j1}} - \frac{1_{(X_j=0)}}{1-P_{j1}}\right)\frac{\partial \vec{P}_{j1}}{\partial\theta}. \tag{33}$$

  Taking the derivative on both sides of equation (33) results in

$$\frac{\partial \frac{\partial\, \vec{\vec{\log}} P_j}{\partial\theta}}{\partial\theta} = \left(\frac{1_{(X_j=1)}}{P_{j1}} - \frac{1_{(X_j=0)}}{1-P_{j1}}\right)\frac{\partial^2 \vec{P}_{j1}}{\partial\theta^2} - \left(\frac{1_{(X_j=1)}}{P_{j1}^2} + \frac{1_{(X_j=0)}}{(1-P_{j1})^2}\right)\frac{\partial \vec{P}_{j1}}{\partial\theta} \otimes \frac{\partial \vec{P}_{j1}}{\partial\theta}. \tag{34}$$

  Note the following expectation:

$$E_j = E(X_j) = P_{j1}, \tag{35}$$

  and

$$\frac{\partial^2\, \vec{\log}\ P_j}{\partial\theta^2} = \frac{\partial \frac{\partial\, \vec{\vec{\log}}\ P_j}{\partial\theta}}{\partial\theta}. \tag{36}$$

  Taking the expectation on both sides of equation (34) results in

$$I_j(\vec{\theta}) = \frac{1}{P_{j1}(1-P_{j1})}\frac{\partial \vec{P}_{j1}}{\partial\theta} \otimes \frac{\partial \vec{P}_{j1}}{\partial\theta}. \tag{37}$$

  With equations (17) and (18), the following is obtained:

$$I_j(\vec{\theta}) = \frac{P_{j1}(1-P_{j1})}{(1+\beta_{3j}e^{(-\vec{\beta}_{2j}\odot\vec{\theta}^T + \beta_{1j})})^2}\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}. \tag{38}$$

  Here, $\otimes$ is a vector cross-product; $\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}$ is a $D \times D$ matrix, and its mth row and nth column element is the product of the mth and nth element of $\vec{\beta}_{2j}$.

- For M-2PPC item j,

$$\frac{\partial\, \vec{\log}\ P_j}{\partial\theta} = \sum_{k=1}^{K_j} 1_{(X_j=k-1)} \frac{\partial\, \vec{\log}\ P_{jk}}{\partial\theta} = \sum_{k=1}^{K_j} 1_{(X_j=k-1)} \frac{\partial \vec{P}_{jk}}{\partial\theta}\frac{1}{P_{jk}}, \tag{39}$$

$$\frac{\partial \frac{\partial\, \vec{\vec{\log}}\ P_j}{\partial\theta}}{\partial\theta} = \sum_{k=1}^{K_j} 1_{(X_j=k-1)}\left(\frac{\partial^2 \vec{P}_{jk}}{\partial\theta^2}\frac{1}{P_{jk}} - \frac{1}{P_{jk}^2}\frac{\partial \vec{P}_{jk}}{\partial\theta} \otimes \frac{\partial \vec{P}_{jk}}{\partial\theta}\right), \tag{40}$$

$$\log P_{jk} = (k-1)\vec{\beta}_{2j} \odot \vec{\theta}^T - \sum_{t=1}^{k}\beta_{\delta_t j} - \log\left(\sum_{m=1}^{K_j} e^{((m-1)\vec{\beta}_{2j}\odot\vec{\theta}^T - \sum_{t=1}^{m}\beta_{\delta_t j})}\right). \tag{41}$$

Taking the derivative on both sides of equation (41) yields the following:

$$\frac{\partial \vec{P}_{jk}}{\partial \theta} = P_{jk}((k-1) - E_j)\vec{\beta}_{2j}. \tag{42}$$

Taking log on both sides and then taking the derivative on both sides of equation (22) results in

$$\frac{\partial \vec{E}_j}{\partial \theta} = \left(\sum_{k=1}^{K_j} (k-1)^2 P_{jk} - E_j^2 \right) \vec{\beta}_{2j} = \sigma_j^2 \vec{\beta}_{2j}, \tag{43}$$

where

$$\sigma_j^2 = \sum_{k=1}^{K_j} (k-1)^2 P_{jk} - E_j^2. \tag{44}$$

Taking the derivative on both sides of equation (42) yields

$$\frac{\partial^2 \vec{P}_{jk}}{\partial \theta^2} = \left(\frac{\partial \vec{P}_{jk}}{\partial \theta}((k-1) - E_j) - P_{jk}\frac{\partial \vec{E}_j}{\partial \theta} \right) \otimes \vec{\beta}_{2j} \tag{45}$$

$$= (P_{jk}((k-1) - E_j)^2 - P_{jk}\sigma_j^2)\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}. \tag{46}$$

Therefore,

$$\frac{\partial \frac{\partial \overrightarrow{\log} P_j}{\partial \theta}}{\partial \theta} = \sum_{k=1}^{K_j} 1_{(X_j = k-1)} \sigma_j^2 \vec{\beta}_{2j} \otimes \vec{\beta}_{2j} = -\sigma_j^2 \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}, \tag{47}$$

and

$$I_j(\vec{\theta}) = \sigma_j^2 \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}. \tag{48}$$

Test Information

Either of the two foregoing definitions for information can be used to obtain test information.

- For the first definition, the test information function at point $\vec{\theta}$ in the direction $\vec{\alpha}$ in θ space can be obtained as

$$I(\vec{\theta})(\vec{\alpha}) = \sum_{j=1}^{J} I_j(\vec{\theta})(\vec{\alpha}). \tag{49}$$

- For the second definition, let

$$I(\vec{\theta}) = \sum_{j=1}^{J} I_j(\vec{\theta}). \tag{50}$$

Assume a unidimensional trait θ_α can be written as

$$\theta_\alpha = \vec{\cos}\,\alpha \odot \vec{\theta}^T. \tag{51}$$

The estimate is

$$\hat{\theta}_\alpha = \vec{\cos}\,\alpha \odot \hat{\vec{\theta}}^T, \tag{52}$$

where $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)$, and $\hat{\vec{\theta}}$ is the estimate of $\vec{\theta}$. The test information function at point $\vec{\theta}$ in θ space in the direction $\vec{\alpha}$ can be obtained by

$$I(\theta_\alpha) = Var(\hat{\theta}_\alpha)^{-1}, \tag{53}$$

where

$$Var(\hat{\theta}_\alpha) = \vec{\cos}\,\alpha \odot Var(\hat{\vec{\theta}}) \odot \vec{\cos}\,\alpha^T, \tag{54}$$

and

$$Var(\hat{\vec{\theta}}) = I(\hat{\vec{\theta}})^{-1}. \tag{55}$$

Note that a necessary condition for the test to be multidimensional is that $I(\hat{\vec{\theta}})$ be positive definite.

Multidimensional Item Parameter Statistics

To be consistent with traditional IRT methods for examining how items are functioning, multidimensional versions of discrimination and difficulty are developed for the M-2PPC model. Multidimensional versions of discrimination and difficulty for the M-3PL model, developed by Reckase (1985) and Reckase and McKinley (1991), are also presented.

Multidimensional Item Discrimination (MDISC)

The multidimensional item discrimination (MDISC) is an item's maximum discrimination in a particular direction of the factor space. MDISC has the same relationship to multidimensional item difficulty as the *a* (discrimination) parameter has to the *b* (difficulty) parameter for the unidimensional IRT model. The MDISC is a measure of an item's capacity to distinguish between examinees who have different locations in the factor space. If an item has a high value of MDISC, then it will provide a relatively large amount of information somewhere in the factor/trait space. MDISC for each item j was defined as the following:

$$\text{MDISC}_j = \left(\sum_{m=1}^{D} \beta_{2jm}^2 \right)^{1/2}. \tag{56}$$

The directional discrimination in the direction $\vec{\alpha}$ is given by

$$\text{MDISC}_{j\alpha} = \sum_{m=1}^{D} \beta_{2jm} \cos \alpha_m, \tag{57}$$

where D is the dimension, and $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)$.

Multidimensional Item Difficulty (MID)

Multidimensional item difficulty (MID) or the signed distance can be interpreted like the difficulty parameter of a unidimensional model. MID for the M-2PL model was initially proposed by Reckase (1985). It represents the distance and direction from the origin in the θ-space to the point of the steepest slope. Two definitions of MID pertaining to the M-2PPC model are developed in this section. To give greater context for these statistics and to derive MID for the M-2PPC model, the meaning of item difficulty for the more familiar 3PL and 2PPC models is also presented.

MID for multiple-choice items. Use MID $_j^D$ to denote the signed distance for item j of dimension D. Using a single dimension, for a 3PL item j, the signed difficulty is defined as the location

$$\mathrm{MID}_j^1 = \frac{\beta_{1j}}{\beta_{2j}}. \tag{58}$$

This definition of MID has the following properties:

- MID is the location, which gives the ability level such that an examinee would have a probability halfway between the distance of guessing and 1 of obtaining the item correctly.
- MID is the location that the item response curve discriminates the most, that is,

$$\frac{\partial E_j^2}{\partial \theta^2} \mid_{\theta=\mathrm{MID}_j^1} = \frac{\partial P_j^2}{\partial \theta^2} \mid_{\theta=\mathrm{MID}_j^1} = 0, \tag{59}$$

  where E_j is the expectation of the jth item.
- For item j,

$$\frac{\beta_{1j} + \log(1 - 2\beta_{3j})}{\beta_{2j}} \tag{60}$$

  is the ability level such that the response curve of 0 and 1 intersects. In the case where the guessing parameter is 0, this intersection is also MID.
- MID is the location that gives the most information, that is,

$$\frac{\partial I_j}{\partial \theta} \mid_{\theta=\mathrm{MID}_j^1} = 0. \tag{61}$$

- For each $j = 1, 2, \ldots, J$, the signed difficulty of an M-3PL item j was extended by Reckase (1985):

$$\mathrm{MID}_j^D = \frac{\beta_{1j}}{\left(\sum_{m=1}^{D} \beta_{2jm}^2\right)^{1/2}}. \tag{62}$$

MID for constructed response items. MID $_{1j}^D$ and MID $_{2j}^D$ are used to denote the two definitions of difficulty for item j of dimension D. For constructed-response items, two ways of defining difficulty are also proposed. In the unidimensional case,

- For item j,

$$\frac{\beta_{\delta_t j}}{\beta_{2j}} \tag{63}$$

  is the ability level such that the response curves $t-1$ and t intersect, $t = 1, 2, \ldots, K_j$. MID in the unidimensional case can be defined as the mean of those intersections, that is,

$$\mathrm{MID}_{1j}^1 = \frac{\sum_{t=1}^{K_j} \beta_{\delta_t j}}{(K_j - 1)\beta_{2j}}. \tag{64}$$

- Another way to define difficulty is the item location at which item information I_j is maximized. That is, the difficulty for item j, MID $_{2j}^1$ is obtained from the following expression:

$$\frac{\partial I_j}{\partial \theta} \mid_{\theta=\mathrm{MID}_{2j}^1} = 0. \tag{65}$$

Note that

$$\frac{\partial I_j}{\partial \theta} = \left[\sum_{k=1}^{K_j} (k-1)^3 P_{jk} - 3E_j \sum_{k=!1}^{K_j} (k-1)^2 P_{jk} + 2E_j^3 \right] \beta_{2j}^3, \tag{66}$$

$$\frac{\partial^2 E_j}{\partial \theta^2} = \left[\sum_{k=1}^{K_j} (k-1)^3 P_{jk} - 3E_j \sum_{k=1}^{K_j} (k-1)^2 P_{jk} + 2E_j^3 \right] \beta_{2j}^2. \tag{67}$$

- Using similar logic, the difficulty for the M-2PPC model for item j can be defined as

$$\mathrm{MID}_{1j}^D = \frac{\sum_{t=1}^{K_j} \beta_{\delta_t j}}{(K_j - 1)(\sum_{m=1}^{D} \beta_{2jm}^2)^{1/2}}. \tag{68}$$

- Once again, another way to define the multidimensional difficulty for the M-2PPC model is the θ such that the item information I_j, with parameter

$$\vec{\beta}_j = (\mathrm{MDISC}_j, \beta_{\delta_2 j}, \ldots, \beta_{\delta_{K_j} j}), \tag{69}$$

is maximized, that is,

$$\frac{\partial I_j}{\partial \theta} \mid_{\theta = \mathrm{MID}_{2j}^D} = 0. \tag{70}$$

The Angle Measure (α)

The final statistic associated with item difficulty is the angle measure. Reckase (1985) proposed describing multidimensional difficulty by both the MID and the angle measure or direction cosines. The direction of greatest slope from the origin with dimension k for item $j = 1, \ldots, J$ is given by

$$\alpha_k = \arccos \frac{\beta_{2jk}}{\sqrt{\sum_{m=1}^{D} \beta_{2jm}^2}}, \tag{71}$$

for $k = 1, 2, \ldots, D$. The angles with the axis will reflect the patterns that are present in the discrimination parameters. Items with high discrimination/factor loadings (e.g., β_{2jk}) on a factor will tend to cluster along that axis. To compare the difficulty of items, they should have similar angle measures.

Analysis of Mixed Format Tests: Using a Writing Assessment

The multidimensional models (i.e., M-3PL and M-2PPC) are applied to a data set containing a mixture of different item types. Data from a state writing assessment consisting of 2,500 Grade 5 students were used. The writing assessment contained 41 multiple-choice and 13 constructed-response items with the following number of maximum score levels: 2, 2, 2, 2, 2, 2, 2, 5, 5, 3, 5, 5, 5. The writing assessment was measured by two standards:

- Standard 1. "Write for a Variety of Purposes." Students write and speak for a variety of purposes and audiences.

- Standard 2. "Write Using Conventions." Students write and speak using conventional grammar, usage, sentence structure, punctuation, capitalization, and spelling.

To demonstrate the calibration of mixed-format tests using the M-3PL and M-2PPC models and examine their factor structure, exploratory type models were implemented. The proposed models varied both in number of dimensions and factor correlations. Moreover, they did not follow a simple structure, characterized by items that are associated only with a particular dimension. Models with one, two, and three dimensions were estimated using three different factor correlations, $r = 0.0, 0.3, 0.7$, for the multidimensional cases, totaling seven calibrations implemented with BMIRT. The choice of correlations is arbitrary, but they also serve to fix the scale. The indeterminacies of the model can be solved by fixing the population parameters of normal with mean 0 and standard deviation of 1 for the unidimensional case, as well as multinormal with mean $(0, 0)$ and specifying the variance and covariance matrix as

$$\sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \tag{72}$$

in the case where $D = 2$.

For each model variation using BMIRT, 15,000 iterations were specified, with the first 3,000 assigned to the "burn-in" phase with arbitrary starting points. Evaluating the MCMC errors using a trace plot showed that the 3,000 iterations allocated for the "burn-in" were sufficient to ensure a stationary chain.

For multiple-choice items (M-3PL), the item parameters are

$$\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j}), \tag{73}$$

with the priors given below.

$$\beta_{1j} \sim N(\mu_{\beta_{1j}}, \sigma^2_{\beta_{1j}}), \tag{74}$$

$$\log(\beta_{2jl}) \sim N(\log(\mu_{\beta_{2j}}), \sigma^2_{\beta_{2j}}), \tag{75}$$

for $l = 1, \ldots, D$.

$$\beta_{3j} \sim beta(a, b), \tag{76}$$

and $a = 100$, $b = 400$, $\mu_{\beta_{2j}} = 1$, $\mu_{\beta_{1j}} = 0$, $\sigma_{\beta_{1j}} = 1$, and $\sigma_{\beta_{2j}} = 1.5$. The priors and proposal functions used gave reasonable acceptance rates in the MCMC sampling.

For constructed-response items (M-2PPC), the item parameters are

$$\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{\delta_2 j}, \ldots, \beta_{\delta_{K_j} j}). \tag{77}$$

The priors are taken to be lognormal for each component of $\vec{\beta}_{2j}$ and normal for $\beta_{\delta_k j}$. The means and standard deviations of the prior distributions are $\mu_{\beta_{2j}} = 1$, $\mu_{\beta_{\delta_k j}} = 0$, $\sigma_{\beta_{\delta_k j}} = 1$, and $\sigma_{\beta_{2j}} = 1.5$, where $k = 2, \ldots, K_j$.

Results From the Exploratory Analysis

Table 1 shows the results for these models using the AIC and the difference chi-square as indicators of model improvement. Because items and ability estimates are performed simultaneously, the degrees of freedom used are reflected in the total number of parameters in Table 1. The

Table 1

Model Fit Statistics Based on One-, Two-, and Three-Dimensional Exploratory Models for the Writing Assessment

Model	Akaike Correlation Criterion	Information Log-Likelihood	χ^2	Items	Number of Parameters Examinees	Total
One dimension	139,079	–66,875		165	2,500	2,665
Two dimensions						
0.0	137,100	–63,331	7,088			
0.3	137,498	–63,530	6,690	219	5,000	5,219
0.7	137,551	–63,556	6,638			
Three dimensions						
0.0	137,870	–61,162	4,338			
0.3	138,600	–61,527	4,006	273	7,500	7,773
0.7	140,283	–62,368	2,376			

development of these fit statistics is given in the appendix. The difference in log-likelihoods multiplied by –2 is asymptotically distributed as a chi-square statistic. The chi-square differences are given, contrasting one dimension with the two-dimensional models and also for the differences between two and three dimensions. The AIC with the lowest value could be designated as the preferred model. The chi-square differences are largest between one and two dimensions compared with the differences between two and three dimensions. Based on AIC and the chi-square differences, the two-dimensional model with $r = 0.0$ provided the best fit to the data, which was slightly better than the two-dimensional model with correlations of 0.3 and 0.7. The two-dimensional parameter estimation with $r = 0.0$ will be used for the demonstration of multidimensional item and test statistics.

Multidimensional Item and Test Statistics

This section presents multidimensional item and test information together with their corresponding statistics (item difficulty and discrimination). To evaluate the characteristics of the multidimensional statistics, their more familiar unidimensional and classical counterparts are included.

Item Information From First Definition

Figures 1 to 4 show multidimensional information as a response surface using the first definition for one multiple-choice Item 17 and three constructed-response Items 45 (two levels), 49 (five levels), and 51 (three levels). They obtain maximum information for angles of about 10, 20, 60, and 50 degrees, respectively. Items 17 and 45 had more information associated with the first dimension (Theta One, θ_1) and Items 49 and 51 relatively more with the second dimension (Theta Two, θ_2). Item 49, which had five levels, had demonstrated high amounts of information associated with many of the two-dimensional score composites. Figure 5 shows item information in the "clamshell" vector format with lines drawn at 10-degree intervals. The individual clamshells identify which direction(s) provide the most information for various areas of the factor plane. It can be seen that information for Item 49 is much greater compared with 17, 45, and 51 and that Item 51 measures many composites of low ability well.

Figure 1
Multidimensional Item Information at Nine Angles, Item 17

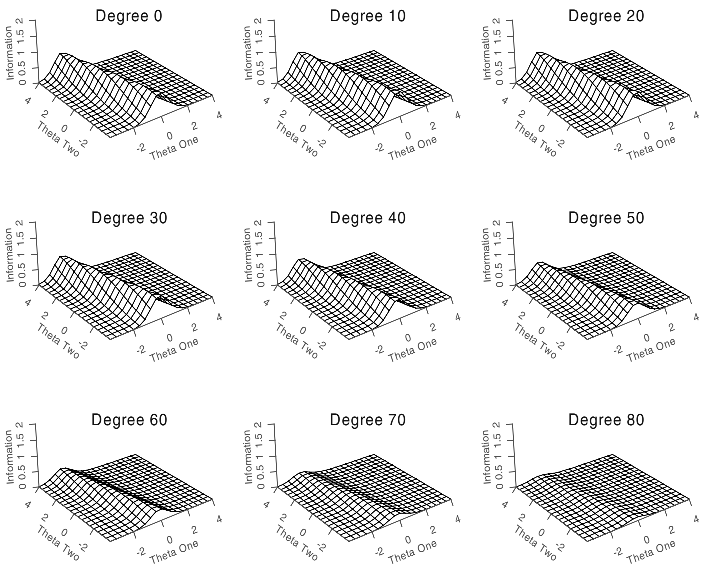
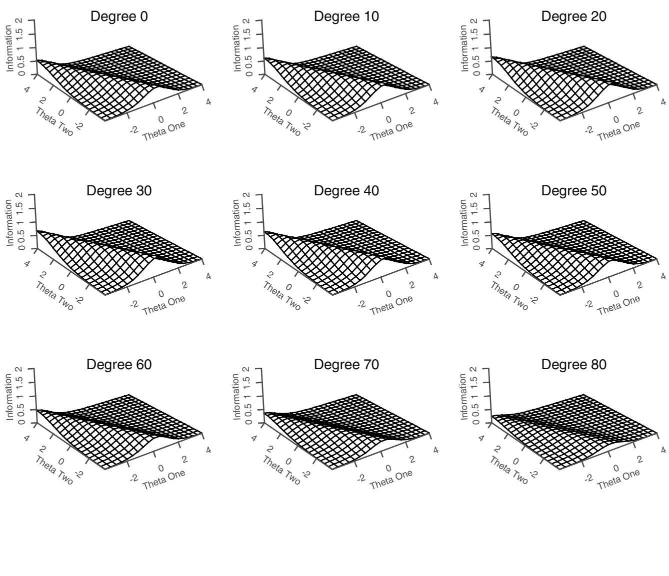


Figure 2
Multidimensional Item Information at Nine Angles, Item 45



Test Information

Figure 6 shows test information using the first definition for angles of 20, 40, 60, and 80 degrees. The most information is given at a 40-degree angle with relatively high amounts shown for both θ_1 and θ_2. Figure 7 shows test information from the second definition for angles of

Figure 3
Multidimensional Item Information at Nine Angles, Item 49

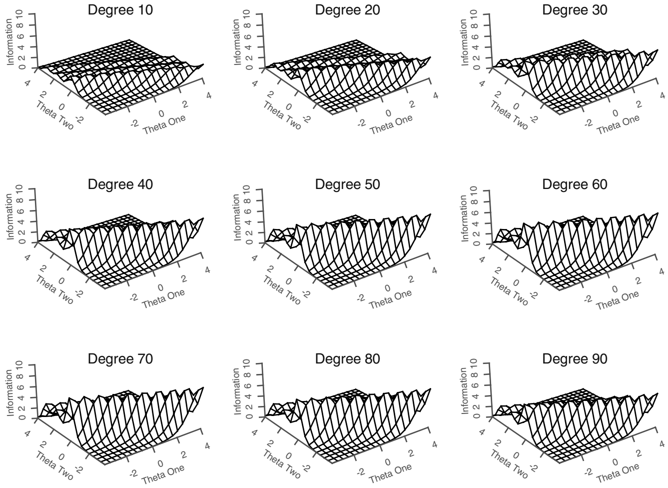
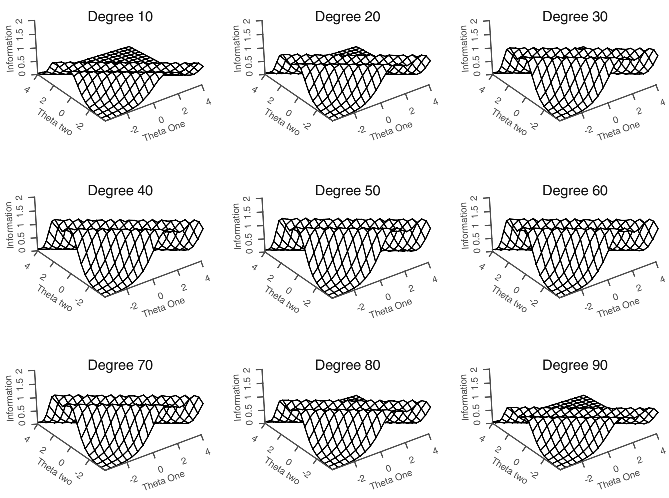


Figure 4
Multidimensional Item Information at Nine Angles, Item 51



10, 15, 20, and 25 degrees. It is clear from using the second definition that more information is associated with θ_1. The best single score will be the composite of the two dimensions with an angle of 15 degrees. As mentioned earlier, BMIRT also estimates examinees' ability. If $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ is the two-dimensional ability estimates for an examinee, then $\hat{\theta} \odot (\vec{\cos} 15^0)^T = \hat{\theta}_1 \cos 15^\circ + \hat{\theta}_2 \cos 75^\circ$ will be the best single-score report for this examinee. Figure 8 displays test information as a "clam-shell" directional vector in θ space for both definitions and essentially replicates the findings from

Figure 5
Multidimensional Item Information as Directional Vectors in θ Space

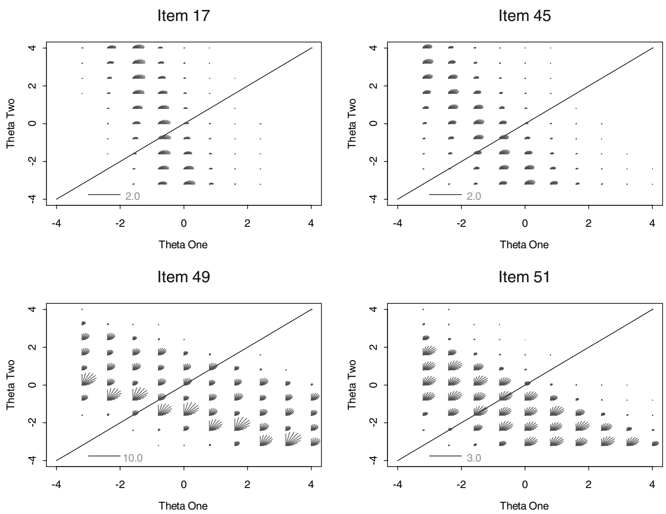
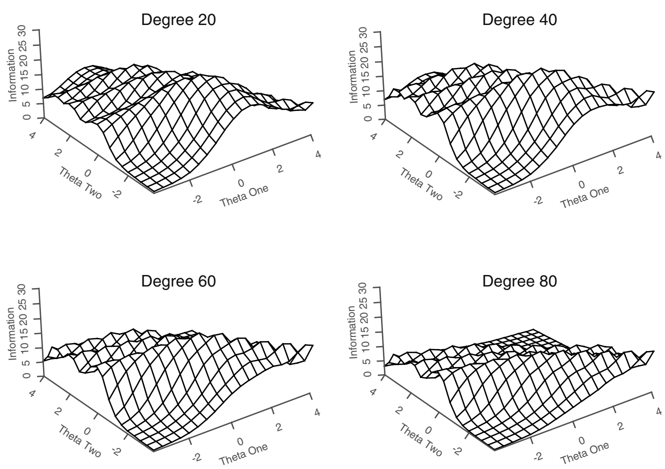


Figure 6
Multidimensional Test Information at Four Angles: Definition 1



surface plots. Multidimensional test information from the first definition shows high amounts of information primarily associated with two-dimensional composites of ability in ranges of $\{(\theta_1, \theta_2) : -1 < \theta_1 < 2, -3 < \theta_2 < 0\}$ and $\{(\theta_1, \theta_2) : -3 < \theta_1 < -1, -1 < \theta_2 < 3\}$. The plot using the second definition also clearly shows that information is primarily associated with θ_1, and the information is higher for $-1 < \theta_1 < 1$.

Figure 7
Multidimensional Test Information at Four Angles: Definition 2

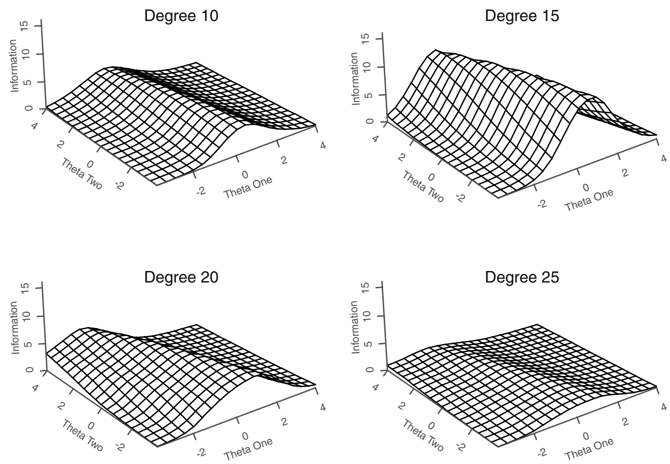
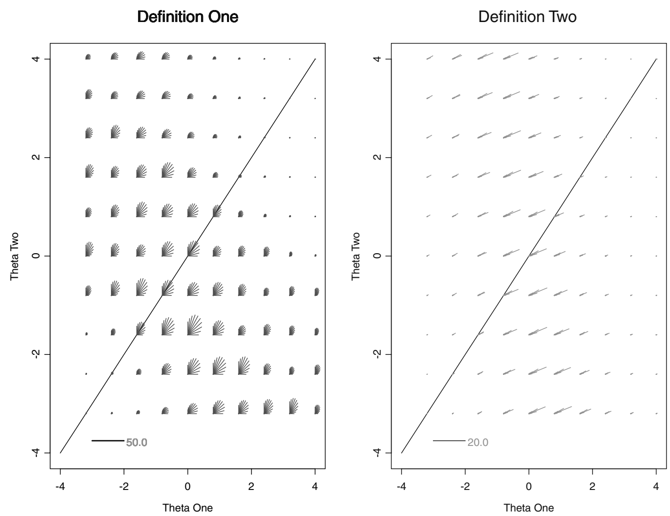


Figure 8
Multidimensional Test Information as Directional Vectors in θ Space



Item Statistics

Table 2 shows the multidimensional, unidimensional, and classical statistics for each item using the one-dimensional and the two-dimensional models with zero correlation between the factors. The first two columns are the estimates for the two discrimination parameters. The 3rd column is the angle measurement, and the 4th column is MDISC. The 5th and 6th columns are the two MID definitions from the two-dimensional model. The two MID definitions for multiple-choice items within a given number

Table 2
Item Statistics

Item		Two Dimensions						One Dimension			Classical	
Item	Level	$\beta_{2,1}$	$\beta_{2,2}$	α_1	MDISC	MID_1^2	MID_2^2	β_2	MID_1^1	MID_2^1	r_{BIS}	p Value
1	1	1.54	0.00	0	1.54	–0.27	–0.27	1.43	–0.31	–0.31	0.45	0.67
2	1	1.25	0.36	16	1.31	–1.22	–1.22	1.28	–1.26	–1.26	0.34	0.81
3	1	1.19	0.38	18	1.25	–1.63	–1.63	1.24	–1.66	–1.66	0.32	0.86
4	1	1.34	0.26	11	1.37	1.22	1.22	1.17	1.30	1.30	0.25	0.35
5	1	2.16	0.84	21	2.32	–2.52	–2.52	2.55	–2.47	–2.47	0.26	0.97
6	1	1.03	0.40	22	1.10	–2.44	–2.44	1.09	–2.46	–2.46	0.26	0.92
7	1	0.55	0.22	22	0.60	0.28	0.28	0.55	0.24	0.24	0.25	0.56
8	1	1.00	0.31	17	1.05	–0.45	–0.45	1.01	–0.50	–0.50	0.38	0.68
9	1	2.55	0.65	14	2.64	0.01	0.01	2.42	–0.01	–0.01	0.52	0.58
10	1	0.99	0.34	19	1.05	–1.01	–1.01	1.04	–1.05	–1.05	0.36	0.79
11	1	1.34	0.29	13	1.37	–0.84	–0.84	1.30	–0.90	–0.90	0.39	0.76
12	1	0.96	0.46	26	1.07	–1.78	–1.78	1.08	–1.79	–1.79	0.32	0.85
13	1	1.11	0.48	24	1.21	–2.05	–2.05	1.21	–2.05	–2.05	0.32	0.90
14	1	1.52	0.42	15	1.58	1.04	1.04	1.43	1.05	1.05	0.32	0.39
15	1	0.87	0.28	18	0.92	–0.21	–0.21	0.90	–0.24	–0.24	0.37	0.63
16	1	1.14	0.26	13	1.17	–1.58	–1.58	1.10	–1.67	–1.67	0.35	0.84
17	1	2.09	0.46	13	2.14	–1.08	–1.08	2.03	–1.12	–1.12	0.45	0.83
18	1	1.68	0.53	18	1.76	–1.45	–1.45	1.76	–1.46	–1.46	0.41	0.88
19	1	1.38	0.55	22	1.48	–0.65	–0.65	1.49	–0.67	–0.67	0.45	0.74
20	1	1.93	0.80	23	2.09	–0.23	–0.23	2.09	–0.26	–0.26	0.51	0.68
21	1	1.51	0.51	19	1.59	–0.03	–0.03	1.57	–0.05	–0.05	0.43	0.61
22	1	1.38	0.57	22	1.50	0.01	0.01	1.48	–0.03	–0.03	0.45	0.62
23	1	1.49	0.41	15	1.55	0.03	0.03	1.48	0.00	0.00	0.42	0.58
24	1	1.69	0.46	15	1.75	–0.50	–0.50	1.68	–0.54	–0.54	0.48	0.70
25	1	1.33	0.30	13	1.37	–0.79	–0.79	1.30	–0.85	–0.85	0.40	0.75
26	1	2.16	0.69	18	2.27	–1.31	–1.31	2.23	–1.33	–1.33	0.45	0.89
27	1	1.26	0.28	13	1.29	–0.88	–0.88	1.23	–0.93	–0.93	0.38	0.79
28	1	1.12	0.27	14	1.15	–0.07	–0.07	1.05	–0.13	–0.13	0.41	0.62
29	1	1.77	0.46	15	1.83	–0.87	–0.87	1.75	–0.91	–0.91	0.51	0.79
30	1	1.71	0.31	10	1.74	–0.41	–0.41	1.55	–0.48	–0.48	0.48	0.71
31	1	2.14	0.40	11	2.18	–0.22	–0.22	1.90	–0.28	–0.28	0.51	0.68
32	1	0.98	0.22	13	1.01	0.16	0.16	0.91	0.13	0.13	0.31	0.58
33	1	0.89	0.33	20	0.95	0.86	0.86	0.91	0.85	0.85	0.32	0.48
34	1	1.49	0.38	14	1.53	–0.33	–0.33	1.44	–0.38	–0.38	0.42	0.66
35	1	1.36	0.51	21	1.46	–0.22	–0.22	1.43	–0.25	–0.25	0.48	0.64
36	1	1.66	0.49	17	1.73	–0.39	–0.39	1.66	–0.44	–0.44	0.49	0.69
37	1	1.25	0.28	13	1.28	–0.47	–0.47	1.19	–0.52	–0.52	0.41	0.71
38	1	1.30	0.53	22	1.41	–1.29	–1.29	1.42	–1.31	–1.31	0.40	0.82
39	1	1.47	0.39	15	1.52	–1.29	–1.29	1.52	–1.31	–1.31	0.39	0.84
40	1	1.45	0.22	9	1.46	1.10	1.10	1.15	1.23	1.23	0.34	0.40
41	1	1.14	0.36	17	1.20	–1.10	–1.10	1.18	–1.14	–1.14	0.38	0.80
42	2	0.97	0.79	39	1.26	–2.30	–4.48	1.19	–2.40	–4.81	0.18	0.99
43	2	0.88	0.31	19	0.94	0.28	0.58	0.92	0.27	0.54	0.38	0.39
44	2	0.60	0.33	29	0.69	–1.23	–2.45	0.69	–1.24	–2.49	0.26	0.83
45	2	1.46	0.75	27	1.64	–0.72	–1.43	1.68	–0.73	–1.46	0.47	0.84

46	2	1.20	0.34	16	1.25	–0.29	–0.56	1.22	–0.30	–0.61	0.46	0.64
47	2	0.37	0.22	31	0.43	0.86	1.80	0.42	0.86	1.73	0.22	0.32
48	2	1.31	0.42	18	1.37	1.01	2.06	1.30	1.04	2.08	0.29	0.10
49	5	1.69	3.51	64	3.90	–0.46	–0.12	1.31	–0.61	0.20	0.70	0.66
50	5	2.23	3.96	61	4.54	–0.51	–0.06	1.67	–0.71	0.30	0.74	0.66
51	3	1.52	1.96	52	2.48	–1.04	–0.87	1.81	–1.20	–2.42	0.58	0.87
52	5	0.94	0.67	35	1.16	–0.37	–1.75	1.16	–0.39	–1.81	0.64	0.61
53	5	0.88	0.61	35	1.07	–0.30	–1.59	1.10	–0.33	–1.65	0.61	0.59
54	5	0.91	0.62	34	1.10	–0.15	–1.25	1.13	–0.17	–1.31	0.63	0.56

Note. MID = multidimensional item difficulty; MDISC = multidimensional item discrimination.

of dimensions are equivalent. The 7th column is the discrimination parameter, β_2, from the unidimensional model. The 8th and 9th columns are the two MID definitions for the unidimensional case. Classical statistics for discrimination and difficulty are given in the last two columns. The 10th column is the correlation between the score for the item and the total score, which is the point-biserial correlation in the case of dichotomously scored items or polyserial correlations for the polytomously scored items. The last column is item p value or percentage of maximum (item mean divided by the range of the item) in the case of polytomously scored items. The two definitions of MID for constructed-response items resulted in values that differed in magnitude. Comparison of MID among items is best for items with relatively similar angle measures. Almost without exception, the multiple-choice items load on the first factor more strongly as well as many of the two-level constructed-response items, as indicated by the angle measure. Several of the constructed-response items load highly on the second factor and have high values of MDISC. For three other constructed-response items (52, 53, 54) with five score levels that tended to load equally on both factors, MDISC was relatively low. A comparison of MDISC with the unidimensional item discriminations β_2 shows that MDISC is almost always larger. However, the observations obtained from this data set might not generalize to other ones.

The correlations among these item statistics are listed in Tables 3 through 5. Some caution is warranted when interpreting these correlations because the relationships among item statistics might be nonlinear. Not surprisingly, $\beta_{2,1}$, the first dimensional discrimination estimates, had a high correlation with β_2, the one-dimensional discrimination estimates, which corresponds with the multiple-choice items loading primarily on the first factor. The second dimensional discrimination, $\beta_{2,2}$, was highly correlated with MDISC, whereas the item-test correlation had the highest

Table 3

Correlations Among Discrimination Statistics

	$\beta_{2,1}$	$\beta_{2,2}$	MDISC	β_2	r_{BIS}
$\beta_{2,1}$	1.00				
$\beta_{2,2}$	0.37	1.00			
MDISC	0.80	0.84	1.00		
β_2	0.93	0.25	0.68	1.00	
r_{BIS}	0.45	0.60	0.64	0.39	1.00

Note. MDISC = multidimensional item discrimination.

Table 4
Correlations Among Difficulty Statistics for Multiple-Choice Items

	MID_1^2	MID_2^2	MID_1^1	MID_2^1	*p* Value
MID_1^2	1.00				
MID_2^2	1.00	1.00			
MID_1^1	0.99	0.99	1.00		
MID_2^1	0.99	0.99	1.00	1.00	
p value	–0.97	–0.97	–0.97	–0.97	1.00

Note. MID = multidimensional item difficulty.

Table 5
Correlations Among Difficulty Statistics for Constructed-Response Items

	MID_1^2	MID_2^2	MID_1^1	MID_2^1	*p* Value
MID_1^2	1.00				
MID_2^2	0.92	1.00			
MID_1^1	0.99	0.90	1.00		
MID_2^1	0.92	0.97	0.90	1.00	
p value	–0.95	–0.85	–0.95	–0.87	1.00

Note. MID = multidimensional item difficulty.

correlation with MDISC and $\beta_{2,2}$. Tables 4 and 5 display the correlations for difficulty statistics for the 3PL and 2PPC items as well as classical ones. These correlations are consistently higher than those given for discrimination in the previous table. Note that MID definitions for multiple-choice items within a given number of dimensions are equivalent. MID^1 (one dimension) and MID^2 (two dimensions) both have high correlations with the *p* values. For constructed-response items, the correlation between the *p* value and the first definition of MID_1 is higher than the correlation between the *p* value and the second definition of MID_2.

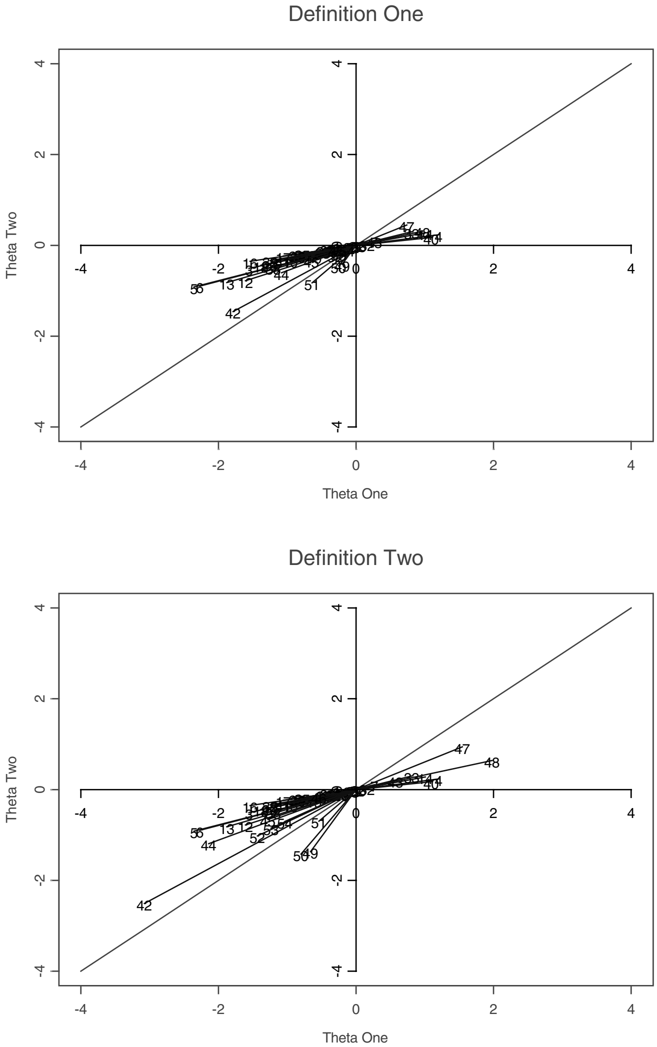
Figure 9 shows item vectors demonstrating both direction and distance obtained from the two-dimensional exploratory model, with a factor correlation of 0. The signed distance for the items in Figure 9 is given for the two definitions of MID (MID_1^2, MID_2^2) accordingly. Because most of the items are relatively easy, the item vectors are clustered in the lower left quadrant, with most of the items positioned at a 15-degree angle.

Discussion

This article presented two multidimensional IRT models using the M-3PL for multiple-choice and the M-2PPC for constructed-response-type tasks that can be applied to tests containing mixed item formats. The multidimensional statistics developed for the M-2PPC model and graphical presentations for the two-dimensional case are necessary to describe item and test functioning for these types of applications. Two definitions of multidimensional test information and difficulties (MID) along with item information were developed here for the M-2PPC model and presented for

Figure 9

Item Locations and Distances Using $\mathrm{MID}_{\{1\}}^{\{2\}}$ (Definition 1) and $\mathrm{MID}_{\{2\}}^{\{2\}}$ (Definition 2)



the M-3PL model, building on the work of Reckase (1985), Reckase and McKinley (1991), and Ackerman (1994, 1996). The results of the exploratory analysis, based on the writing assessment, indicated that the models and derived statistics functioned as expected when compared with their unidimensional and classical counterparts and the intercorrelations.

There were clearly differences between the two definitions for test information with respect to the amount of information displayed. However, there is utility in the use of the first definition of information because it gives an approximation for item information. The ‘‘clamshell’’ plots show information in a more compact format, which can be easily inspected to determine the composite of abilities that a test measures best. Surface plots for items display information when a more granular view is necessary. Although it was not presented here, it might be of interest to compare the relative contribution of multidimensional information provided collectively by multiple-choice and constructed-response item types. MDISC was low for several constructed-response items, with five score levels, which is likely due to essentially equal loading on both factors. The values obtained for the two definitions of MID for constructed-response items varied, but the correlations between them shown in Table 5 were very high. Either definition of MID could be used, but the first one, defined in equation (64) by the intersection of the response curves, is closer to the original conception of item location exemplified in the unidimensional partial credit model. The angle measure and loadings for multiple-choice items were uniformly associated with the first factor, which was not the case for the constructed-response tasks.

Various confirmatory models could be investigated by associating items with specific dimensions or traits. One prevalent hypothesis is that multiple-choice and constructed-response items measure different traits or dimensions (Traub, 1993). Factors could be hypothesized to form either on the basis of item format (multiple choice or constructed response) or along the two content standards, "Write for a Variety of Purposes" and "Write Using Conventions." The findings of the dimensional structures from the exploratory calibration did not quite form factors solely based on the two content standards or strictly according to the item types, which is not unusual in many applications. No definitive answer can be arrived at here using this limited application about the dimensionality of different item types. This study's conjecture is that the skills and knowledge assessed by an item (i.e., its content) may be just as important as the item's format type. A confirmatory analysis that associated dimensions with content objectives and estimated subscale scores based on those objectives can be found in Yao and Boughton (2005a). Although these interpretations of the dimensions and analysis are of importance to practice, they are not addressed here. The models and statistics developed provide the tools necessary to further explore these types of questions within an IRT framework.

Appendix

Computation of Model Fit Statistics

Let

$$G_k^2 = -2\log(P(X\mid\theta,\beta)), \tag{78}$$

$$df_k = (k+2)J_1 + \sum_{j=1}^{J_2} L(j) + (k-1)J_2 + k(N-D-D^2), \tag{79}$$

$$f_{k+1} = df_{k+1} - df_k, \tag{80}$$

$$d_{k+1} = G_k^2 - G_{k+1}^2, \tag{81}$$

for $k = 1, 2, \ldots, D$, and D is the number of dimensions. Note that in BMIRT program, the item, examinees, and population parameters are all estimated together. Therefore, all these parameters are included in the degrees of freedom. Here,

- $L(j)$ is the level of the jth constructed response item.
- J_1 is the number of multiple-choice items.

- J_2 is the number of constructed items.
- N is the number of examinees.
- df_k is the number of parameters for dimension k.
- d_k is the difference chi-square when adding dimension k. The contribution of dimension k added to the model can be judged significant if the corresponding difference chi-square is statistically significant.
- f_k is the difference in the degrees of freedom.

Let

$$AIC_k = G_k^2 + 2df_k. \tag{82}$$

AIC $_k$ or difference chi-square can be used to judge the fit of a k-dimensional IRT model.

Note

1. $\vec{a} \odot \vec{b}^T = \sum_{i=1}^{n} a_i b_i$, where $\vec{a} = (a_1, \ldots, a_{n)}$, and $\vec{b} = (b_1, \ldots, b_n)$.

References

Ackerman, T. (1989). Unidimensional IRT calibration of compensatory and noncompensatory items. *Applied Psychological Measurement, 13,* 113-127.

Ackerman, T. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18,* 257-275.

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311-329.

Ackerman, T., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational psychological tests. *Educational Measurement: Issues and Practice, 22,* 37-51.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 62,* 317-332.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement, 49,* 789-805.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Educational and Psychological Measurement, 16,* 327-343.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 3,* 137-154.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

Green, B. F. (1990). *Notes on the item information function in the multidimensional compensatory* IRT *model* (Report No. 88-10). Baltimore: John Hopkins University, Psychometric Laboratory.

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement, 20,* 155-168.

Lane, S. (2005, April). *Status and future directions for performance assessments in education.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Muraki, E. (1999). POLYFACT Version 2 [Computer program]. Princeton, NJ: Educational Testing Service.

Patz, R., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342-346.

Patz, R., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

Patz, R., & Yao, L. (2003, April). *Hierarchical and multidimensional models for vertical scaling.* Paper presented at the annual meeting of National Council on Measurement in Education, Chicago.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Traub, R. E. (1993). On the equivalence of traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum.

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

Waller, N. (2002). MicroFACT: A microcomputer factor analysis program for ordered polytomous data and mainframe size problems [Computer software and manual]. St. Paul, MN: Assessment Systems Corporation.

Wilson, D., Wood, R., & Gibbons, R. D. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer software]. Mooresville, IN: Scientific Software.

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. (2004a, August). *Bayesian multivariate item response theory and BMIRT software.* Paper presented at the Joint Statistical Meetings, Toronto, Canada.

Yao, L. (2004b). LinkMIRT: Linking of multivariate item response model [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L., & Boughton K. (2005a). *A multidimensional item response theory approach for improving subscore proficiency estimation in cognitive diagnostic assessments.* Manuscript under review.

Yao, L., & Boughton, K. A. (2005b). *Multidimensional parameter recovery: Markov chain Monte Carlo versus NOHARM.* Manuscript under review.

Acknowledgments

The authors express their gratitude to Rich Patz in supporting the development of BMIRT and to the editor and anonymous reviewers for their comments and suggestions.

Author's Address

Address correspondence to Lihua Yao, CTB/McGraw-Hill, 20 Ryan Ranch Rd, Monterey, CA 93940; e-mail: Lihua_Yao@ctb.com.