The Multidimensional Random Coefficients Multinomial Logit Model

Raymond J. Adams, Australian Council for Educational Research

Mark Wilson and Wen-chung Wang, University of California, Berkeley

A multidimensional Rasch-type item response model, the multidimensional random coefficients multinomial logit model, is presented as an extension to the Adams & Wilson (1996) random coefficients multinomial logit model. The model is developed in a form that permits generalization to the multidimensional case of a wide class of Rasch models, including the simple logistic model, Masters' partial credit model, Wilson's ordered partition model, and Fischer's linear logistic model. Moreover, the model includes several existing multidimensional models as special cases, including Whitely's multicomponent latent trait model, Andersen's multidimensional Rasch model for repeated testing, and Embretson's multidimensional Rasch model for learning and change. Marginal maximum likelihood estimators for the model are derived and the estimation is examined using a simulation study. Implications and applications of the model are discussed and an example is given. *Index terms: EM algorithm, marginal maximum likelihood, multidimensional IRT models, multinomial logit models, Rasch models.*

A basic assumption of most item response models is that the items measure one common latent trait (Hambleton & Murray, 1983; Lord, 1980). This is the unidimensionality assumption. There are, however, at least two reasons why this assumption can be problematic. First, researchers have argued that the unidimensionality assumption is inappropriate for many standardized tests that are deliberately constructed from subcomponents that are assumed to measure different traits (Ansley & Forsyth, 1985). Although it is often argued that item response models are robust to such violations of unidimensionality, particularly when the traits are highly correlated, this is not always the case. For example, in computerized adaptive testing, in which examinees are administered different combinations of test items, trait estimates may reflect different composites of traits underlying performance and thus cannot be compared directly (Way, Ansley, & Forsyth, 1988). Further, when a test contains mutually exclusive subsets of items or when the underlying dimensions are not highly correlated, the use of a unidimensional model can bias parameter estimation, adaptive item selection, and trait estimation (Folk & Green, 1989).

Second, and perhaps more importantly, the demands of current assessment practice often go beyond unidimensional summaries of examinee traits or achievement. Modern practice often requires the examination of single performances from multiple perspectives; for example, it may be useful to code an examinee's performance not only for its accuracy or correctness but also for the strategy used in the performance or the conceptual understanding displayed by the performance.

The potential usefulness of multidimensional item response theory (MIRT) models has been recognized for many years. Recently, there has been considerable work on the development of MIRT models and, in particular, on the consequences of applying unidimensional models to real or simulated multidimensional data (e.g., Ackerman, 1992; Andersen, 1985; Camilli, 1992; Embretson, 1991; Folk & Green, 1989; Glas, 1992; Kelderman & Rijkes, 1994; Luecht & Miller, 1992;

Applied Psychological Measurement, Vol. 21 No. 1, March 1997, 1–23

Oshima & Miller, 1992; Reckase, 1985; Reckase & McKinley, 1991). Despite this, the application of MIRT models in practical testing situations has been limited. This has probably been due to the statistical problems that have been involved in fitting such models and in the difficulty associated with the interpretation of the parameters of existing MIRT models.

Two problems have not been adequately addressed in the research literature on MIRT models. First, the research has focused on dichotomously scored items; therefore, most of the existing models and computer programs cannot be applied to multidimensional polytomously scored items, such as those that generally arise in performance assessment. Second, the limited flexibility of existing models and computer programs does not match the complexity of real testing situations, which may involve complexities such as raters and item sampling.

The purpose of this research was to provide some solutions to these problems. A multidimensional Rasch-type model, called the multidimensional random coefficients multinomial logit model (MRCMLM), several subclasses of the model, and an estimation algorithm for the MRCMLM are described. Simulation studies were conducted to examine the accuracy of both item and population parameter estimation and the adequacy of asymptotic standard error estimates. A real dataset of an educational application of the MRCMLM was also analyzed.

The Unidimensional Random Coefficients Multinomial Logit Model

The MRCMLM is a multidimensional extension of the unidimensional random coefficients multinomial logit model (RCMLM; Adams & Wilson, 1996). A brief explanation of the RCMLM is useful for understanding the MRCMLM.

The RCMLM and examples of its use have been described in Wilson & Adams (1995) and Wilson & Wang (in press). The RCMLM is a generalized Rasch model that integrates many existing Rasch models, such as the simple logistic model (Rasch 1960), the linear logistic latent trait model (Fischer, 1973), the rating scale model (Andrich, 1978), the partial credit model (PCM; Masters, 1982), and the ordered partition model (Wilson, 1992). In addition, the RCMLM provides substantial flexibility in allowing the design of customized models for particular test situations.

The RCMLM is also equivalent to, or a modest extension of, some other extended Rasch models that impose linear structures on the item parameters. These include Linacre's (1989) FACETS model, Glas & Verhelst's (1989) extension to the PCM (Glas, 1989), and Fischer & Pononcy's (1994) extended rating scale and PCMs for assessing change.

Assume that n items are indexed $i = 1, \ldots, n$ with each item consisting of $K_i + 1$ response alternatives ($k = 0, 1, \ldots, K_i$). Use the vector valued random variable, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iK_i})'$, where

$$X_{ik} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } k \\ 0 & \text{otherwise} \end{cases} , \tag{1}$$

to indicate the $K_i + 1$ possible responses to item i.

A response in category 0 is denoted by a vector of 0s. This effectively makes the 0 category a reference category and is necessary for model identification. The choice of this as the reference category is arbitrary and does not affect the generality of the model. The $\mathbf{X}_i$ can be collected into the single vector $\mathbf{X}' = (\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_n)$ called the response vector (or pattern). Realizations of each of these random variables are indicated by their lower case equivalents: $\mathbf{x}$, $\mathbf{x}_i$, and x_{ik}.

The items are described by a vector $\boldsymbol{\xi}' = (\xi_1, \xi_2, \ldots, \xi_p)$ of p parameters. Linear combinations of these are used in the response probability model to describe the empirical characteristics of the response categories of each item. These linear combinations are defined by design vectors $\mathbf{a}_{ik}$ ($i = 1, \ldots, n$ and $k = 1, \ldots, K_i$), each of length p, that can be collected to form a design matrix $\mathbf{A} = (\mathbf{a}_{11}, \mathbf{a}_{12}, \ldots, \mathbf{a}_{1K_1}, \mathbf{a}_{21}, \ldots, \mathbf{a}_{2K_2}, \ldots, \mathbf{a}_{1K_n})$. Adopting a very general approach to the definition of items, in conjunction with the imposition of a linear model on the item parameters, allows a general model to be written that includes the wide class of existing Rasch models mentioned above and allows the development of new types of Rasch models [e.g., the item bundle models of Wilson & Adams (1995)].

An additional feature of the RCMLM is the introduction of a scoring function that allows the specification of the score or performance level that is assigned to each possible item response. To do this, the notion of a response score b_{ik} is introduced that gives the performance level of an observed response in category k of item i. The b_{ik} can be collected in a vector as $\mathbf{b}' = (b_{11}, b_{12}, \ldots, b_{1K_1}, b_{21}, b_{22}, \ldots, b_{2K_2}, \ldots, b_{1K_n}$. (By definition, the score for a response in the 0 category is 0, but other responses may also be scored 0.)

In the majority of Rasch model formulations, there has been a one-to-one match between the category to which a response belongs and the score that is allocated to the response. In the simple logistic model, for example, it has been standard practice to use the labels 0 and 1 to indicate both the categories of performance and the scores. A similar practice has been followed with the rating scale and PCMs, in which each different possible response is seen as indicating a different level of performance so that the category indicators 0, 1, 2, and so forth that are used serve both as scores and labels. The use of $\mathbf{b}$ as a scoring function allows a more flexible relationship between the qualitative aspects of a response and the level of performance that it reflects. Scoring functions of this type are available in Wilson's (1992) ordered partition model, Muraki's (1992) generalized PCM, and the LOGIMO model of Kelderman & Rijkes (1994).

Letting θ be the latent variable, the RCMLM item response probability model can be written as

$$P(\mathbf{X}_{ik} = 1; \mathbf{A}, \mathbf{b}, \boldsymbol{\xi} | \theta) = \frac{\exp\left(b_{ik}\theta + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)}{\sum_{k=1}^{K_i} \exp\left(b_{ik}\theta + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)} \tag{2}$$

and a response vector probability model as

$$P(\mathbf{X} = \mathbf{x} | \theta) = \Psi(\theta, \boldsymbol{\xi}) \exp\left[\mathbf{x}'(\mathbf{b}\theta + \mathbf{A}\boldsymbol{\xi})\right] \;, \tag{3}$$

with

$$\Psi(\theta, \boldsymbol{\xi}) = \left\{ \sum_{z \in \Omega} \exp\left[\mathbf{z}'(\mathbf{b}\theta + \mathbf{A}\boldsymbol{\xi})\right] \right\}^{-1} \;, \tag{4}$$

where $\boldsymbol{\Omega}$ is the set of all possible response vectors.

The Multidimensional Random Coefficients Multinomial Logit Model

The MRCMLM is an extension of the RCMLM that assumes that a set of D traits underlie the persons' responses. The D latent traits define a D-dimensional latent space, and the persons' positions in the D-dimensional latent space are represented by the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)$. The scoring function of response category k in item i now corresponds to a $D \times 1$ column vector rather than a scalar, as in the RCMLM. A response in category k on Dimension $d (d = 1, \ldots, D)$ of item i is scored b_{ikd}. The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik} = (b_{ik1}, b_{ik2}, \ldots, b_{ikD})'$, then be collected into the scoring submatrix for item i, $\mathbf{B}_i = (\mathbf{b}_{i1}, \mathbf{b}_{i2}, \ldots, \mathbf{b}_{iD})'$, and then collected into a scoring matrix $\mathbf{B} = (\mathbf{B}'_1, \mathbf{B}'_2, \ldots, \mathbf{B}'_n)'$ for the entire test. If the item parameter vector, $\boldsymbol{\xi}$, and the design matrix, $\mathbf{A}$, are defined as they were in the RCMLM, the probability of a response in category k of item i is modeled as

$$P(\mathbf{X}_{ik} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) = \frac{\exp\left(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)}{\sum_{k=1}^{K_i} \exp\left(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)} \quad . \tag{5}$$

For a response vector,

$$f(\mathbf{x}; \boldsymbol{\xi} | \boldsymbol{\theta}) = \Psi(\boldsymbol{\theta}, \boldsymbol{\xi}) \exp\left[\mathbf{x}'(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\xi})\right] \quad , \tag{6}$$

with

$$\Psi(\boldsymbol{\theta}, \boldsymbol{\xi}) = \left\{ \sum_{z \in \Omega} \exp\left[\mathbf{z}'(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\xi})\right] \right\}^{-1} \quad . \tag{7}$$

The difference between the RCMLM and the MRCMLM is that the trait parameter is a scalar, θ, in the former, and a $D \times 1$ column vector, $\boldsymbol{\theta}$, in the latter. Likewise, the scoring function of response k to item i is a scalar, b_{ik}, in the former, whereas it is a $D \times 1$ column vector, $\mathbf{b}_{ik}$, in the latter.

An item with four response categories and design matrices,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{8}$$

and

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{9}$$

is modeled as follows. Substituting these matrices into Equation 5 gives

$$\begin{aligned} P(\mathbf{X}_{11} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) &= \exp(\theta_1 + \xi_1)/D \\ P(\mathbf{X}_{12} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) &= \exp(\theta_1 + \theta_2 + \xi_1 + \xi_2)/D \\ P(\mathbf{X}_{13} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) &= \exp(\theta_1 + \theta_2 + \theta_3 + \xi_1 + \xi_2 + \xi_3)/D \quad , \end{aligned} \tag{10}$$

where

$$D = 1 + \exp(\theta_1 + \xi_1) + \exp(\theta_1 + \theta_2 + \xi_1 + \xi_2) + \exp(\theta_1 + \theta_2 + \theta_3 + \xi_1 + \xi_2 + \xi_3) \quad , \tag{11}$$

which is a multidimensional PCM (note that the constraints that would be necessary for the identification of this model have not been imposed). This example is discussed again below.

As an extension of the unidimensional RCMLM, the MRCMLM inherits all of the flexibility of the RCMLM. Importantly, this allows the specification of a range of multidimensional models by imposing linear constraints on the item parameters. For example, multidimensional forms of models such as the rating scale models and linear logistic text model can all be easily specified.

Estimation of the MRCMLM

Marginal Maximum Likelihood Estimation

The MML approach alleviates the problem of inconsistent estimates, which result from joint maximum likelihood estimation, by assuming that persons have $\boldsymbol{\theta}$ vectors that are sampled from a population in which the distribution of $\boldsymbol{\theta}$ is given by the multivariate density function $g(\boldsymbol{\theta};\boldsymbol{\alpha})$ (Bock & Aitkin, 1981). $G(\boldsymbol{\theta};\boldsymbol{\alpha})$ is the corresponding distribution function, and $\boldsymbol{\alpha}$ indicates a vector of parameters that characterize the distribution. Here, two alternative forms of g are considered—(1) the multivariate normal $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the population mean and standard deviation, respectively, and $\boldsymbol{\alpha} = (\boldsymbol{\mu},\boldsymbol{\Sigma})$; and (2) a step distribution defined on a prespecified set of nodes. For the step distribution, a fixed vector of Q points, $(\Theta_{d1},\Theta_{d2},\dots,\Theta_{dQ})$, was selected for each latent dimension, and the nodes of each of the $r = 1,\dots,Q^D$ cartesian coordinates are defined as

$$\boldsymbol{\Theta}_r = \left(\Theta_{1k_1},\Theta_{2k_2},\dots,\Theta_{Dk_D}\right) \quad \text{with } k_1 = 1,\dots,Q;\ k_2 = 1,\dots,Q;\ \dots\ ;\ k_D = 1,\dots,Q\,. \tag{12}$$

The parameters that characterize the distribution are $\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\dots,\boldsymbol{\Theta}_{Q^D}$ and $w_1,w_2,\dots,w_{Q^D}$, where $g(\boldsymbol{\Theta};\boldsymbol{\alpha}) = w_r$ is the density at node r. The grid points are specified a priori and the density values are estimated. Although formally this model uses a discrete trait distribution on the set of prespecified nodes, it is probably best viewed as an opportunity to approximate an arbitrary continuous trait distribution.

To construct the marginal likelihood, Equation 6 (the conditional response vector probability) is combined with the multivariate density function $g(\boldsymbol{\theta};\boldsymbol{\alpha})$ to obtain the marginal density of the response pattern $\mathbf{x}_j$ of person j,

$$f\left(\mathbf{x}_j;\boldsymbol{\xi}\right) = \int_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta},\boldsymbol{\xi}) \exp\left[\mathbf{x}_j'\left(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\xi}\right)\right] dG(\boldsymbol{\theta};\boldsymbol{\alpha})\ . \tag{13}$$

Then the likelihood for a set of N response patterns is

$$\Lambda\left(\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{X}\right) = \prod_{j=1}^{N} \int_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta},\boldsymbol{\xi}) \exp\left[\mathbf{x}_j'\left(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\xi}\right)\right] dG(\boldsymbol{\theta};\boldsymbol{\alpha})\ . \tag{14}$$

It follows that the likelihood equations for the item parameters are

$$\begin{aligned}\frac{\partial \log\Lambda\left(\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{X}\right)}{\partial\boldsymbol{\xi}} &= \sum_{j=1}^{N} \int_{\boldsymbol{\theta}} \frac{\partial \log f\left(\mathbf{x}_j;\boldsymbol{\xi},\boldsymbol{\alpha}\right)}{\partial\boldsymbol{\xi}} dH\left(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{x}_j\right) \\ &= \mathbf{A}' \sum_{j=1}^{N} \left[\mathbf{x}_j - \int_{\boldsymbol{\theta}} \mathrm{E}_x(\mathbf{z}|\boldsymbol{\theta}) dH\left(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{x}_j\right)\right] = \mathbf{0}\ ,\end{aligned} \tag{15}$$

where

$$\mathrm{E}_x(\mathbf{z}|\boldsymbol{\theta}) = \Psi(\boldsymbol{\theta},\boldsymbol{\xi}) \sum_{z\in\Omega} \mathbf{z}\exp\left[\mathbf{z}'\left(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\xi}\right)\right] \tag{16}$$

is the expected response pattern, and $H(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{x}_j)$ is the distribution function of the marginal density of $\boldsymbol{\theta}$ given $\mathbf{x}_j$, the density of which is given by

$$h\left(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{x}_j\right) = \frac{f\left(\mathbf{x}_j;\boldsymbol{\xi}|\boldsymbol{\theta}\right) g\left(\boldsymbol{\theta};\boldsymbol{\alpha}\right)}{f\left(\mathbf{x}_j;\boldsymbol{\xi}\right)}\ . \tag{17}$$

For the population parameters, the likelihood equations depend on the choice of the density function, g. In the multivariate normal case,

$$\frac{\partial \log\Lambda\left(\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X}\right)}{\partial\boldsymbol{\mu}} = \sum_{j=1}^{N} \int_{\boldsymbol{\theta}} \frac{\partial \log g(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial\boldsymbol{\mu}} dH\left(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{x}_j\right) = \mathbf{0} \tag{18}$$

gives

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{\theta}} \boldsymbol{\theta} dH(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{x}_j) \ , \tag{19}$$

and

$$\frac{\partial \log \Lambda(\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X})}{\partial \boldsymbol{\Sigma}} = \sum_{j=1}^{N}\int_{\boldsymbol{\theta}} \frac{\partial \log g(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} dH(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{x}_j) = \mathbf{0} \tag{20}$$

gives

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{\theta}} (\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})' dH(\boldsymbol{\theta};\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{x}_j) \ . \tag{21}$$

In the step distribution case, the likelihood equations for the weights can be written as

$$\frac{\partial \log \Lambda(\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{X})}{\partial \mathbf{w}} = \sum_{j=1}^{N}\sum_{q=1}^{Q} \frac{\partial \log w_q}{\partial \mathbf{w}} h(\boldsymbol{\Theta}_q;\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{x}_j) = \mathbf{0} \ . \tag{22}$$

It can be shown that

$$w_q = \frac{1}{N}\sum_{j=1}^{N} h(\boldsymbol{\Theta}_q;\boldsymbol{\xi},\boldsymbol{\alpha}|\mathbf{x}_j) \quad \text{for } q = 1,\ldots,Q-1 \ . \tag{23}$$

For both the normal case (likelihood Equations 15, 19, and 21) and the step distribution case (likelihood Equations 15 and 23), Bock & Aitkin's (1981) formulation of the EM algorithm (Dempster, Laird, & Rubin, 1977) can be used to estimate the model parameters:

Step 1. Calculate the marginal posterior density of $\boldsymbol{\theta}$ given x_n at iteration t using

$$h\left[\boldsymbol{\theta};\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}|\mathbf{x}_j\right] = \frac{f\left[\mathbf{x}_j;\hat{\boldsymbol{\xi}}^{(t)}|\boldsymbol{\theta}\right] g\left[\boldsymbol{\theta};\hat{\boldsymbol{\alpha}}^{(t)}\right]}{f\left[\mathbf{x}_j;\hat{\boldsymbol{\xi}}^{(t)}\right]} \ , \tag{24}$$

where $\hat{\boldsymbol{\xi}}^{(t)}$ and $\hat{\boldsymbol{\alpha}}^{(t)}$ are estimates of $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$ at iteration t, respectively.

Step 2. Using the Newton-Raphson method, solve

$$\mathbf{A}'\sum_{j=1}^{N}\left\{\mathbf{x}_j - \int_{\boldsymbol{\theta}} \mathrm{E}_z\left[\mathbf{z};\hat{\boldsymbol{\xi}}^{(t+1)}|\boldsymbol{\theta}\right] dH\left[\boldsymbol{\theta};\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}|\mathbf{x}_j\right]\right\} = \mathbf{0} \ . \tag{25}$$

to produce estimates of $\hat{\boldsymbol{\xi}}^{(t+1)}$.

Step 3. In the normal case, produce direct estimates of $\hat{\boldsymbol{\mu}}^{(t+1)}$ and $\hat{\boldsymbol{\Sigma}}^{(t+1)}$ using

$$\hat{\boldsymbol{\mu}}^{(t+1)} = \frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{\theta}} \boldsymbol{\theta} dH\left[\boldsymbol{\theta};\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}|\mathbf{x}_j\right] \tag{26}$$

and

$$\hat{\boldsymbol{\Sigma}}^{(t+1)} = \frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{\theta}} \left[\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}^{(t+1)}\right]\left[\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}^{(t+1)}\right]' dH\left[\boldsymbol{\theta};\boldsymbol{\xi}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}|\mathbf{x}_j\right] . \tag{27}$$

In the step distribution case, produce estimates of $\hat{\mathbf{w}}^{(t+1)}$ using

$$\hat{\mathbf{w}}_q^{(t+1)} = \frac{1}{N}\sum_{j=1}^{N} h\left[\boldsymbol{\Theta}_q;\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}|\mathbf{x}_j\right] . \tag{28}$$

To produce initial values, $\hat{\boldsymbol{\xi}}^{(0)}$ and $\hat{\boldsymbol{\alpha}}^{(0)}$, proceed as follows. Let f_{ik} be the probability of a person with $\theta = 0$ responding in category k on item i; then from Equation 2 it follows that

$$\log\left(\frac{f_{ik}}{f_{i0}}\right) = \mathbf{a}'_{ik}\boldsymbol{\xi} \quad \text{for} \quad k = 1,\ldots,K_i . \tag{29}$$

Now if p_{ik} is the observed proportion of persons responding in category k on item i, calculate

$$c_{ik} = \log\left(\frac{p_{ik}}{p_{i0}}\right) \quad \text{for} \quad k = 1,\ldots,K_i , \tag{30}$$

gather the c_{ik} in the vector $\mathbf{C} = (c_{11}, c_{12}, \ldots, c_{nK_n})'$, and then solve $\mathbf{C} = \mathbf{A}\boldsymbol{\xi}$ to get $\hat{\boldsymbol{\xi}}^{(0)} = \mathbf{A}^{-}\mathbf{C}$, where $\mathbf{A}^{-}$ is a generalized inverse of $\mathbf{A}$.

Although a similar approach can be used to produce initial values for $\boldsymbol{\alpha}$ from estimates of persons' latent vectors, in the normal case it is just as convenient to begin with the assumption that $\hat{\boldsymbol{\mu}}^{(0)} = \mathbf{0}$ and $\hat{\boldsymbol{\Sigma}}^{(0)} = \mathbf{I}$, and for the step distribution case use initial weights that are proportional to the multivariate normal density with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$.

In the normal case, both Steps 2 and 3 of the algorithm require an integration over the marginal posterior; in the step distribution case, the density is discrete and the integral becomes a summation. In the normal case this integration cannot be obtained in a closed form and must be approximated. Because of its simplicity and flexibility [again following Bock & Aitkin (1981)], a quadrature approximation is used. Using the same grid of Q^D points described for the step distribution, and deriving appropriate weights from the multivariate normal $\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, use the following approximations in Steps 2 and 3:

$$\int_{\boldsymbol{\theta}} \mathrm{E}_z\left[\mathbf{z};\hat{\boldsymbol{\xi}}^{(t+1)}\middle|\boldsymbol{\theta}\right] dH\left[\boldsymbol{\theta};\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}\middle|\mathbf{x}_j\right] \approx \sum_{r=1}^{Q^D} \mathrm{E}_z\left[\mathbf{z};\hat{\boldsymbol{\xi}}^{(t+1)}\middle|\boldsymbol{\Theta}_r\right] h\left[\boldsymbol{\Theta}_r;\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}\middle|\mathbf{x}_j\right] , \tag{31}$$

$$\int_{\boldsymbol{\theta}} \boldsymbol{\theta}\, dH\left[\boldsymbol{\theta};\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}\middle|\mathbf{x}_j\right] \approx \sum_{r=1}^{Q^D} \boldsymbol{\Theta}_r h\left[\boldsymbol{\Theta}_r;\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}\middle|\mathbf{x}_j\right] , \tag{32}$$

and

$$\int_{\boldsymbol{\theta}} \left[\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}^{(t+1)}\right]\left[\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}^{(t+1)}\right]' dH\left[\boldsymbol{\theta};\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}\middle|\mathbf{x}_j\right]$$

$$\approx \sum_{r=1}^{Q^D} \left[\boldsymbol{\Theta}_r - \hat{\boldsymbol{\mu}}^{(t+1)}\right]\left[\boldsymbol{\Theta}_r - \hat{\boldsymbol{\mu}}^{(t+1)}\right]' h\left[\boldsymbol{\Theta}_r;\hat{\boldsymbol{\xi}}^{(t)},\hat{\boldsymbol{\alpha}}^{(t)}\middle|\mathbf{x}_j\right] . \tag{33}$$

Although this approach is both straightforward and flexible, it is computationally expensive because the range of the summations increases exponentially with the dimensionality of the problem. In practice, this approach can only reasonably address problems of three or four dimensions. Its great advantage is that it allows almost total freedom in the choice of the population density, g. [For the multivariate normal population density, the model can be estimated with the computer program MATS (Wu, Adams, & Wilson, 1995). For information about this program, contact the first author. For the step distribution, a more rudimentary program called MRCMLM can be used. For further information, contact the third author.]

Conditional Maximum Likelihood Estimation

Because the MRCMLM is a Rasch model, it is possible to obtain consistent estimates of $\boldsymbol{\xi}$ using conditional maximum likelihood (CML) estimation (Andersen, 1970). For CML estimation, a two-step procedure like that proposed by Anderson & Madsen (1977) and implemented by Hoijtink (1995) could be used.

Anderson & Madsen (1977) factored the likelihood (Equation 13) into two components—one dependent on the item parameters only and a second that depends on both the item and population parameters [see equation (11.5) in Hoijtink (1995)]. Maximizing the first of these two components produces CML item parameter estimates. These CML estimates then can be substituted into the second component of the likelihood, which can be maximized to produce estimates of the population parameters. The primary advantage of CML is that it yields consistent item parameter estimates regardless of the correctness of the specification of the population distributions, G (Pfanzagl, 1994). This is in contrast to the MML approach, which produces biased item parameter estimates if G is misspecified (Zwinderman, 1991). In CML, the population parameters that are estimated in the second step are estimates of the parameters of a misspecified population distribution, even though consistent estimates of the item parameters are obtained.

The primary disadvantage of CML, and the reason it was not used here, is that it can only be used with Rasch models. If an alternative model, such as Muraki's (1992) generalized PCM is posited as the measurement model, then CML estimation is not possible. An additional, albeit minor, disadvantage is that using CML item parameter estimates in a second step as if they were the "true" item parameters has an as yet to be investigated influence on the estimation of the population parameters.

For a nonparametric population distribution, CML is asymptotically efficient (Pfanzagl, 1994). de Leeuw & Verhelst (1986) showed that CML and MML are asymptotically equivalent. For the parametric approach (e.g., normal), Pfanzagl (1994) also showed that when using the dichotomous Rasch model, MML is asymptotically efficient but CML will be asymptotically inefficient. Clearly, the examination of the relative strengths and weaknesses of CML and MML estimation in applications of the MRCMLM is a fertile area for future research.

Identification of the Multidimensional Model

The model in Equation 13 will not generally be identified unless certain constraints are placed on the design matrices **A** and **B**. In addressing the problem of identification, the goal is to determine the conditions that must be satisfied by **A** and **B** to ensure that if

$$\mathbf{x}^{\mathrm{T}}(\mathbf{B}\boldsymbol{\mu}+\mathbf{A}\boldsymbol{\xi}) = \mathbf{x}^{\mathrm{T}}(\mathbf{B}\boldsymbol{\mu}^{*}+\mathbf{A}\boldsymbol{\xi}^{*}) \tag{34}$$

for every possible response vector $\mathbf{x}$, then $\boldsymbol{\mu} \equiv \boldsymbol{\mu}^{*}$ and $\boldsymbol{\xi} \equiv \boldsymbol{\xi}^{*}$. For the normal case, Volodin & Adams (1995) showed that the following are necessary and sufficient conditions for the identification of Equation 6. These conditions can be shown to hold for the nonparametric case, which is also considered here.

Proposition 1. If D is the number of latent dimensions, P is the length of the parameter vector, $\boldsymbol{\xi}$, $K_i + 1$ is the number of response categories for item i, and $K = \sum_{i=1}^{n} K_i$, then Equation 14 can only be identified if $P + D \leq K$.

Proposition 2. If D is the number of latent dimensions and P is the length of the parameter vector, $\boldsymbol{\xi}$, then Equation 14 can only be identified if $\text{rank}(\mathbf{A}) = P$, $\text{rank}(\mathbf{B}) = D$, and $\text{rank}([\mathbf{B} + \mathbf{A}]) = P + D$.

Proposition 3. If D is the number of latent dimensions, P is the length of the parameter vector, $\boldsymbol{\xi}$, $K_i + 1$ is the number of response categories for item i, and $K = \sum_{i=1}^{n} K_i$, then Equation 14 can be identified if and only if $\text{rank}([\mathbf{B} + \mathbf{A}]) = P + D \leq K$.

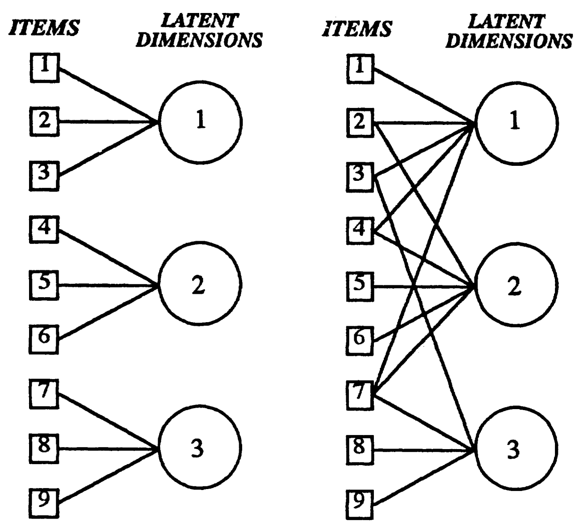
Subclasses of the MRCMLM

To assist in the discussion of different types of multidimensional models and tests, the notions of within- item and between-item multidimensionality are introduced. A test is regarded as multidimensional between items if it consists of several unidimensional subscales. A test is considered multidimensional within items if each of the items relates to more than one latent dimension. The distinction between the within-item and between-item multidimensional models is shown in Figure 1.

Figure 1a depicts a between-item multidimensional test that consists of nine items measuring three latent dimensions. Figure 1b depicts a within-item multidimensional test with nine items and three dimensions. If all of the items in Figure 1 were dichotomous, then the matrices

Figure 1
Multidimensionality at the Item Level

a. Within-Item b. Between-Item



$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{35}$$

and

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{36}$$

would give the MRCMLM for the between-item multidimensional test, and

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{37}$$

and

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{38}$$

would give the MRCMLM for the within-item multidimensional test.

Multidimensional Between-Item Models

Tests that contain several subscales each measuring related, but supposedly distinct, latent dimensions are very commonly encountered in practice. In such tests, each item belongs to only one particular subscale and there are no items in common across the subscales. In the past, item response modeling of such tests has proceeded by either applying a unidimensional model to each of the scales separately or by ignoring the multidimensionality and treating the test as unidimensional. Both of these methods have weaknesses that make them less desirable than undertaking a joint, multidimensional, calibration.

For example, consider the SAT (Donlon, 1984), which contains verbal and quantitative components, but each component is comprised of a distinct set of items. To apply this approach, the SAT is treated as a unidimensional test and each individual is considered to have a single ability because the distinction between quantitative and verbal abilities is lost. In practice this assumption is not particularly harmful to the item parameter estimates if, as is often the case, the two dimensions are highly correlated; but it clearly is of concern if the underlying dimensions are not strongly related. Regardless of the magnitude of the correlation, this approach has a theoretical impurity—fitting a unidimensional model to a test that is multidimensional—that makes it unsatisfactory. Further, the "invisibility" of the subscales means that it is not possible to examine relationships between the dimensions.

Under the second approach, which Davey & Hirsch (1991) call the *consecutive* approach, the test is analyzed separately for each dimension. This approach is often preferred to joint unidimensional analysis because it recognizes the multidimensionality of the test and allows for separate measurements on each dimension. Although it is possible to examine the relationships between the separately measured latent dimensions, such analyses must take the measurement error associated with the dimensions into account, particularly when the subscales have a small number of items. Subscales of at least 20 items are probably desirable before the measurement error can be ignored safely.

A shortcoming of the consecutive approach is its failure to use all available data. In practice it is common for the subscales to be measuring distinct, but highly correlated dimensions; if this is the case, then a joint analysis with an appropriate multidimensional model will lead to improved estimation of the item parameters and ability predictions. The advantage of a model like the MRCMLM with data of this type is that (1) it explicitly recognizes the test developer's intended structure, (2) it provides direct estimates of the relations between the latent dimensions, and (3) it draws on the (often strong) relationship between the latent dimensions to produce more accurate parameter estimates and individual measurements.

In the absence of an accessible and fully operational multidimensional model, the procedures used in the National Assessment of Educational Progress (NAEP) (see Beaton, 1987) can be used as an alternative to a model like the MRCMLM. In the NAEP approach, separate unidimensional item response models are fit to each subscale for the purpose of estimating item parameters only. In the second stage, these parameters are held fixed and the parameters of the underlying multivariate latent space are estimated. In general, the use of a MIRT model such as the MRCMLM is preferred over the NAEP procedure for the following reasons: (1) a MIRT model is a single rather than a multistep analysis; (2) it provides better estimates of the item parameters [although some experience suggests that the gain may be small (Adams, Wilson & Wu, in press)]; and (3) unlike the NAEP procedure, it can be applied to tests that have within-item multidimensionality.

Multidimensional Within-Item Models

If the set of items in a test measures more than one latent dimension and some of the items require abilities from more than one dimension, then the test is termed within-item multidimensional. Three kinds of situations in which within-item multidimensional models can arise are distinguished. First, there is the situation described above in which item developers create items that require abilities that correspond to more than one of the subscales of a multiscale test.

The second kind of within-item multidimensionality occurs when a single performance by an individual (e.g., an essay) is judged on two dimensions, such as theme understanding and writing ability. Each individual is assigned two results for the items—one for theme understanding and the other for writing ability. If the assignment of the results on the two dimensions are independent (conditional on θ), then for modeling purposes it can be considered as a single item with as many outcomes as there are different response patterns to the two ratings, or as two separate items. Consider the following example in which, for the sake of simplicity only, it is assumed that the item is dichotomously scored (0 and 1) on both dimensions. Then the possible responses are (0,0), (0,1), (1,0), (1,1) (recall that the first category is omitted because it is a reference category used for identification). In the notation used here, this can be expressed as one item using the design and scoring submatrices

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{39}$$

and

$$\mathbf{B}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{40}$$

or as two items with the design and scoring submatrices

$$\mathbf{A}_1 = \mathbf{A}_2 = [1] \ , \tag{41}$$

$$\mathbf{B}_1 = [0 \quad 1] \ , \tag{42}$$

and

$$\mathbf{B}_2 = [1 \quad 0] \ . \tag{43}$$

It can be shown that Equations 39–40 and 41–43 will result in the same MRCMLM. In general, however, the first approach is more flexible because it generalizes to more complicated situations, such as the third type of within-item multidimensionality.

The third kind of within-item multidimensionality deals with complications such as those in which some response patterns are not theoretically possible, or in which different achievement levels require different latent abilities. In the former case, consider an item in which three performance levels are identified and that can be solved using one of two strategies; however, one performance level is not attainable with a particular choice of strategy. Under these circumstances, it is not possible to use the second of the above methods because there is dependence between the two types of item outcomes.

In the latter case (different achievement levels require different latent abilities), consider the item in Figure 2, which was used by Masters (1982) and Wright & Masters (1982), as an example of a partial credit item. In applying a unidimensional model to this item, Masters assumed that the

Figure 2
A Partial Credit Item

$$\sqrt{9.0/_{0.3} - 5} = ?$$

0	*No steps taken*
1	$9.0/_{0.3} = 30$
2	$30 - 5 = 25$
3	$\sqrt{25} = 5$

same ability dimension is required for the solution of each step. The necessary design and scoring submatrices for the MRCMLM are

$$\mathbf{A}_i = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{44}$$

and

$$\mathbf{B}_i = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} . \tag{45}$$

In some contexts, it might be argued that the three steps needed to solve this problem require different latent abilities. If this is the case, a suitable multidimensional PCM can be derived in the same way that Masters derived the unidimensional PCM (e.g., Masters, 1988), and it can be fit with the MRCMLM using the following design and scoring submatrices

$$\mathbf{A}_i = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{46}$$

and

$$\mathbf{B}_i = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} . \tag{47}$$

The Relationship Between the MRCMLM and Some Multidimensional Rasch Models

Multidimensional Rasch models have been previously proposed by a number of authors, including Rasch (1961), Whitely (1980), Andersen (1985), and Embretson (1991). Each of these models can be shown to be a special case of the MRCMLM. Whitely's multicomponent latent trait model can be formulated as a MRCMLM for a within-item multidimensional test of the type shown in Equations 37 and 38. Andersen's model for repeated testings [and its extension, the multidimensional Rasch model for learning and change (Embretson, 1991)] can be formulated as a between-item multidimensional MRCMLM with **A** used to impose equality constraints on some of the item parameters.

The expression of the MRCMLM given in Equation 5 disguises the fact that the MRCMLM can be used in contexts in which either a compensatory or a noncompensatory item response model is considered desirable. Although each of the examples provided above are compensatory, the MRCMLM can also be used in a noncompensatory context; for example, consider Whitely's (1984) multicomponent model for independent components. Under this model, the success on a single item, i, requires success on T independent components (the item is made up of subcomponents and an individual must succeed on all subcomponents before success on the whole item is recorded), and the probability of success on the item is given by

$$P(\mathbf{X}_{n1} = 1; \boldsymbol{\xi} | \boldsymbol{\theta}) = \prod_{t=1}^{T} \frac{\exp(\theta_t + \xi_t)}{1 + \exp(\theta_t + \xi_t)} , \tag{48}$$

where θ_t is the ability on the tth component, and ξ_t is a difficulty component for the tth component. If the product in Equation 48 is expanded,

$$P(\mathbf{X}_{n1} = 1; \boldsymbol{\xi} | \mathbf{u}) = \frac{\exp\left(\sum_{t=1}^{T} \theta_t + \sum_{t=1}^{T} \xi_t\right)}{\prod_{t=1}^{T} [1 + \exp(\theta_t + \xi_t)]} . \tag{49}$$

In this form, the multicomponent model for independent components can be shown to be a special case of the MRCMLM. Item i with T components is treated as a single item with 2^T possible response categories, and the denominator is the sum over the set of numerators that arises when the components are modeled as separate items. Wilson & Adams (1995) illustrated a range of applications of this type for the unidimensional version of the RCMLM.

Simulation Studies

Simulation studies were conducted to investigate the properties of the MRCMLM and the estimation method proposed here. Because the MRCMLM is a general model, it was not possible to vary all conditions in the simulation study. Thus, two simulation studies were conducted—a nonparametric study (i.e., no distributional assumptions were made) and a parametric study (a normally distributed population was assumed). The first simulation study focused on the item parameter estimates; the second considered the population parameter estimates and the accuracy of asymptotic estimates of error variance as well as the item parameter estimates.

Method

Design. In the nonparametric case, six factors were manipulated and two levels of each of these factors were selected: (1) the number of latent dimensions (two and three), (2) model class (multidimensional between-item and multidimensional within-item), (3) model type (dichotomous and partial credit), (4) number of items (10 per dimension and 20 per dimension), (5) sample size (N = 300 and N = 700), and (6) the magnitude of the latent correlation (0.0 and .5). 10 replications of each of the 64 resulting combinations were run.

For both simulation studies, a normal population model was used for the data generation. Also, the EM algorithm was terminated when the largest absolute change in any parameter estimate became less than .005. There appears to be little guidance in the literature about the appropriate choice of a stopping rule for the EM algorithm. Given the slow rate of convergence of the EM algorithm, particularly when it is close to the solution, perhaps a stopping rule that is much smaller than .005 would be preferable. Experience shows (e.g., Wang, 1994; Wu & Adams, 1993) that selecting a value less than .005 does not markedly improve the properties of the estimators, nor does it appear to influence any reasonable substantive interpretations that might be based on the parameter estimates. This issue does, however, warrant further investigation.

All numerical integrations performed in both simulation studies used a fixed grid of 10 uniformly spaced nodes per dimension; that is, $Q = 10$. This resulted in a grid of 100 points for the two-dimensional models (2DMs), and 1,000 points for the three-dimensional models (3DMs). For unidimensional item response models, the appropriate choice of Q has been considered: BILOG's (Mislevy & Bock, 1983) default value of Q is 10 for tests of less than 50 items and 20 for tests of more than 50 items; Wilson & Adams (1993) recommended between 10 and 20 nodes; Wu & Adams (1993) recommend 8 to 15 nodes; and TESTFACT (Wilson, Wood, & Gibbons, 1991) uses a default of 10 for a one-dimension model, 5 for two-dimension models, and 3 for three dimensions or more.

Data generation. For both simulation studies, data were generated as follows:

Step 1. A set of item parameters was specified and remained fixed for all data generation.

Step 2. A random sample of $\boldsymbol{\theta}$ vectors of appropriate size was drawn from an assumed multivariate normal population.

Step 3. The known fixed item parameters and the random latent vectors $\boldsymbol{\theta}$ were then combined to calculate the probability of each response category in each item for each simulated examinee. These probabilities were then compared to a uniform (0,1) random number to allocate responses to specific categories.

The random number generators used in Steps 2 and 3 were seeded differently for each of the 10 replications within each of the 64 cells of the design.

Results

Item parameter estimation with nonparametric methods. For each of the 640 replications, bias statistics (i.e., the parameter estimate minus the generating value) were calculated for all item parameters. The proportion of those statistics that were significantly biased was computed by using a z test with a two-tailed .05 significance level. The percentage of item parameters significantly biased at the .05 level is shown in Table 1, along with the number of replicated parameter estimates that were computed for each cell (R). The overall percentage was 9.08 for the 2DMs and 11.60 for the 3DMs. Therefore, the item parameters for the 2DM and the 3DM were marginally biased, with those for the 3DMs somewhat more biased.

To illustrate the magnitude of bias in each cell of the design, the following two indexes were used for each of the 64 conditions. If ζ is the generating value of any parameter in the model, and $\hat{\zeta}_k$ is its estimate in the kth replication, then for each of the conditions,

$$\text{BIAS}(\zeta) = \frac{1}{10}\sum_{k=1}^{10}\left(\zeta - \hat{\zeta}_k\right) \tag{50}$$

and

$$\text{RMSE}(\zeta) = \left[\frac{1}{10}\sum_{k=1}^{10}\left(\zeta - \hat{\zeta}_k\right)^2\right]^{1/2} \tag{51}$$

Table 1
Percentage of Item Parameter Estimates Significantly Biased and Number of Replicated Parameter Estimates (R) for the Nonparametric Simulation Study

	2DMs		3DMs	
Factor and Condition	Bias	R	Bias	R
Model Class				
Between-Item Models	7.44	7,200	9.78	10,800
Within-Item Models	10.72	7,200	13.42	10,800
Model Type				
Dichotomous Models	3.89	4,800	8.88	7,200
Partial Credit Models	14.26	9,600	14.32	14,400
Number of Items				
10 Per Dimension	7.95	4,800	10.83	7,200
20 Per Dimension	10.21	9,600	12.37	14,400
Sample Size				
$N = 300$	7.70	7,200	8.61	10,800
$N = 700$	10.46	7,200	14.59	10,800
Latent Correlation				
$r = .5$	9.07	7,200	9.80	10,800
$r = 0.0$	9.08	7,200	13.40	10,800
Overall	9.08	14,400	11.60	21,600

The bias and root mean square error (RMSE) indexes are shown in Figures 3 and 4 for the 2DM and 3DM. In each figure, the lower pair of lines show the mean absolute bias across all item parameter estimates and the 10 replications for each of the 64 cells in the design for correlations of 0.0 and .05. The upper pairs of lines similarly show the RMSEs. In both figures there is evidence that the RMSEs and biases were larger for the within-item multidimensional models (Figures 3b, 3d, 4b, and 4d) than for the between-item multidimensional models (Figures 3, 3c, 4a, and 4c). Further, the discrepancy was larger for the 3DMs (Figure 4) than for the 2DMs (Figure 3).

Although Table 1 indicates that more of the bias values were significant than would be expected by chance, Figures 3 and 4 show that the magnitude of that bias was small when compared to the sampling variation in the estimates. Although work still needs to be done on this probable bias, it is clear that its magnitude is such that it is unlikely to affect substantive interpretations that result from applications of the MRCMLM.

Parameter estimation with parametric methods. Table 2 shows the results for the parametric simulation. The item response model used in this simulation was a two-dimensional between-item PCM with a latent correlation of .5. Table 2 shows the generating parameters of all item and population parameters, the mean of their estimates across replications, bias, RMSEs, the means of the estimated asymptotic standard errors, the sampling variance in the estimation, and the ratio of the sampling variance to the asymptotic standard error.

For the item parameters, the discrepancies between the generating values and the mean estimated values was small in substantive terms, but large enough to warrant further investigation—this result was consistent with the findings for the nonparametric case. For the population parameters, the discrepancies were negligible. The ratios of the asymptotic standard errors to the sampling errors were close to 1, which suggests that the asymptotic error estimates provided a good approximation to the sampling variation.

Conclusions

The above results provide support for the feasibility of the MRCMLM and the quadrature approach to estimation. Wang (1994) reported a more extensive range of simulations that further

Figure 3

Bias and RMSE for the Item Parameters of Two-Dimensional Dichotomous Models and PCMs, for Short and Long Tests and N = 300 and N = 700

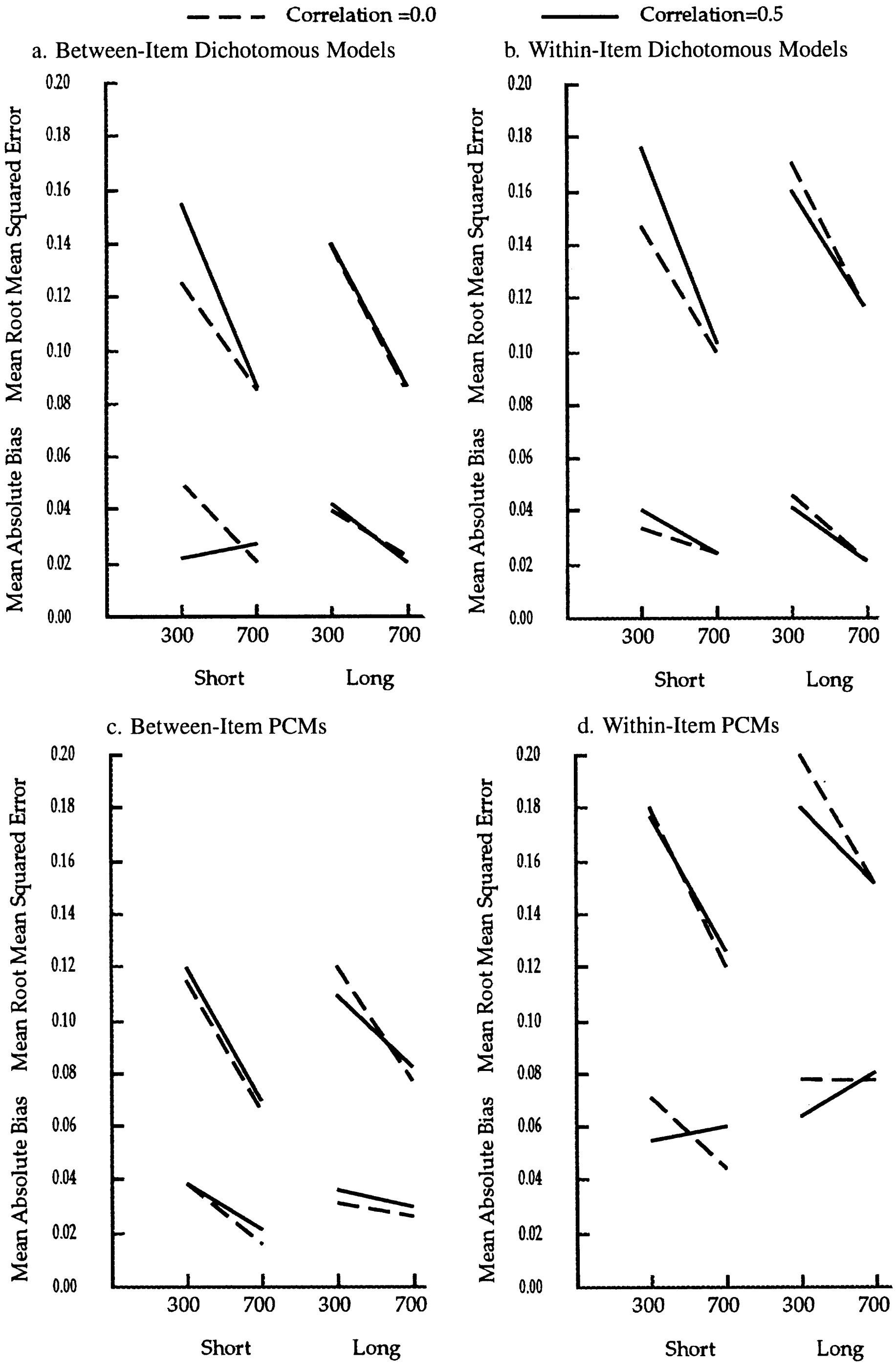


Figure 4
Bias and RMSE for the Item Parameters of Three-Dimensional Dichotomous Models and PCMs, for Short and Long Tests and $N = 300$ and $N = 700$

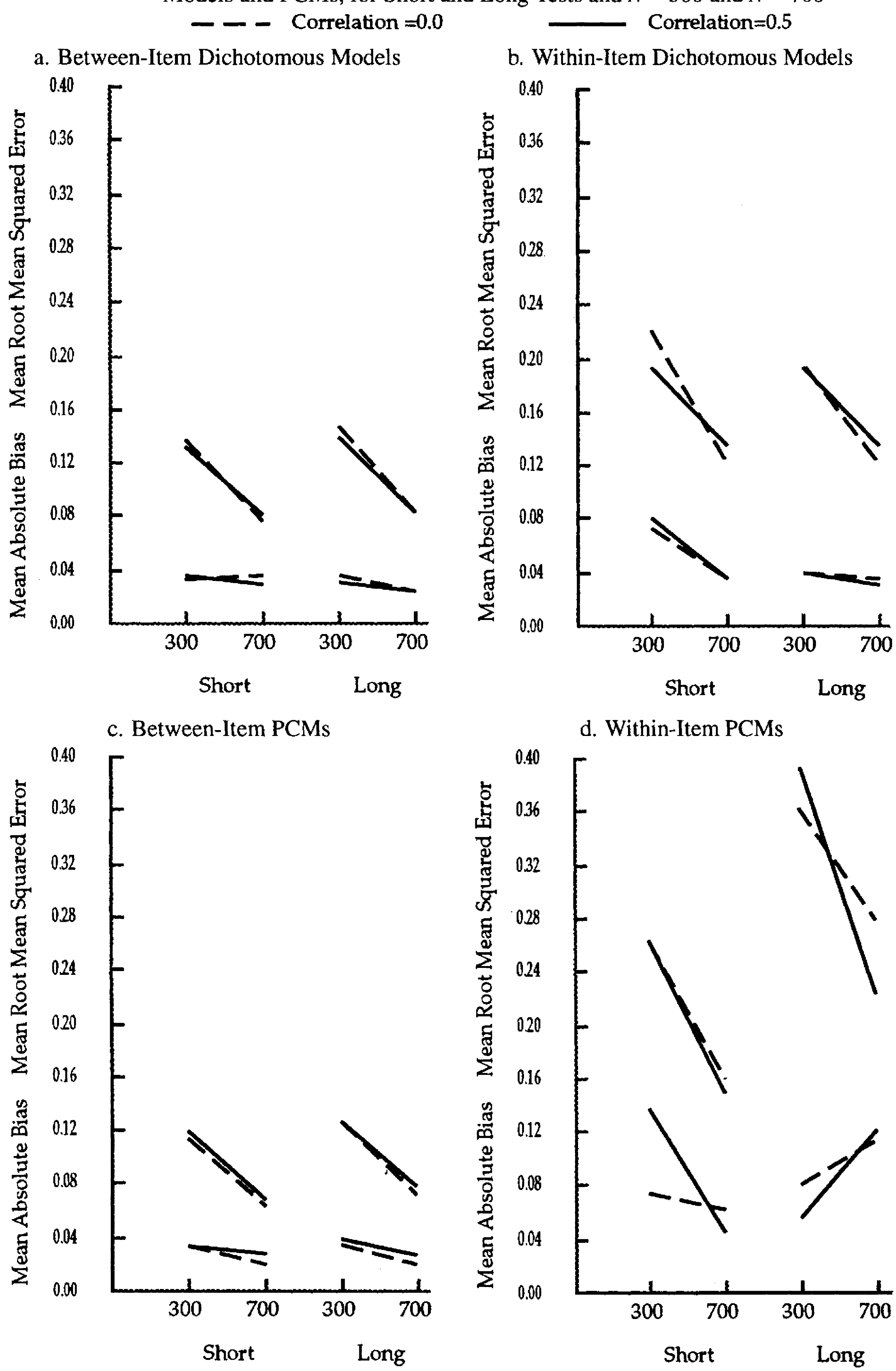


Table 2
Parameter Recovery and the Ratio of Estimated Standard Errors (SEs) to Sampling Variation for the Two-Dimensional Between-Item PCM for the Parametric Simulation Study

Parameter	Generating Values	Mean of Estimates	Bias	RMSE	Mean of Asymptotic SE	Sampling Variation	Ratio
δ_1	−1.2	−1.24294	−.04294	.06259	.047922	.04554	1.05228
δ_2	.3	.35168	.05168	.10605	.088882	.09261	.95974
δ_3	−.6	−.58974	.01026	.05142	.053356	.05039	1.05887
δ_4	.4	.44634	.04634	.07947	.063842	.06456	.98888
δ_5	0.0	.01328	.01328	.07144	.066088	.07020	.94143
δ_6	.6	.60117	.00117	.08352	.082147	.08352	.98356
δ_7	.6	.63616	.03616	.05136	.039901	.03647	1.09409
δ_8	.8	.83584	.03584	.09422	.090169	.08714	1.03476
δ_9	.7	.71432	.01432	.08213	.083317	.08087	1.03026
δ_{10}	−1.0	−.98646	.01354	.05867	.057201	.05709	1.00195
δ_{11}	.4	.43470	.03470	.07181	.060599	.06287	.96387
δ_{12}	−.5	−.51678	−.01678	.07604	.072068	.07416	.97179
δ_{13}	.8	.75589	−.04411	.08491	.070541	.07256	.97218
δ_{14}	0.0	−.00632	−.00632	.05232	.052241	.05194	1.00579
δ_{15}	.9	.90155	.00155	.09582	.095503	.09581	.99680
δ_{16}	.5	.50859	.00859	.03787	.036482	.03688	.98921
δ_{17}	.6	.62554	.02554	.07106	.069829	.06631	1.05307
δ_{18}	.7	.69061	−.00939	.10715	.093023	.10674	.87149
μ_1	0.0	−.00291	−.00291	.04979	.04740	.04970	.95371
μ_8	0.0	−.00139	−.00139	.04822	.04724	.04820	.97996
σ_{11}	1.0	.97735	−.02265	.08629	.09682	.08326	1.16286
σ_{28}	1.0	.99318	−.00682	.08742	.09740	.08715	1.11764
σ_{18}	.5	.49050	−.00950	.06262	.05740	.06190	.92744

support the major conclusions found here:

1. There may be a small amount of bias in the estimation of item parameters for both the normal and the nonparametric models. This bias warrants further investigation, but it is small both in substantive terms and when compared to typical sampling variation.
2. When a normal population distribution is assumed, the parameters of that distribution are well estimated.
3. Asymptotic standard error estimates provide an adequate indication of the sampling variation in the estimators.

Example Application

Instrument

This example is based on the Structure of the Learning Outcome (SOLO) Taxonomy (Biggs & Collis, 1982), which is a method of classifying learner responses to test items according to the structure of the response elements into five levels: prestructural, unistructural, multistructural, relational, and extended abstract. A mathematical problem-solving test containing seven separate stimuli was used (Wilson, 1992). Each stimulus is accompanied by three open-ended, dichotomously scored, items. The first item assessed the unistructural response element, the second assessed the multistructural response element, and the third the relational response element. Because of the age of the examinees (grades 4, 6, and 8), the extended abstract response element was not assessed.

Analysis

The levels of the SOLO taxonomy were viewed as dimensions, so that the seven bundles of three items were viewed as three subscales of seven items each. These data were analyzed three ways, repeating the comparison of Wilson (1988), and extending that comparison by including the MRCMLM. First, the data were treated as unidimensional and a standard Rasch simple logistic model was fit to all 21 items. The unidimensional model (UM) required the estimation of 22 parameters—20 item parameters and the population mean and variance. Second, a consecutive analysis was conducted by fitting a unidimensional Rasch model to each subscale separately. The consecutive model (CM) had parameters for the mean and variance of each dimension and six item parameters for each of the three seven-item subscales. One item parameter on each subscale serves as an identification constraint. Third, a between-item MRCMLM was fit to the data that included 18 item parameters, three means, three variances, and three covariances [the multidimensional model (MM)].

Under the UM, each individual was regarded as having a single ability, whereas under the CM and MM each individual was regarded as having three distinct abilities. Under the CM, correlations between the latent dimensions were obtained by first estimating abilities for individuals on each dimension [expected a posteriori (EAP) estimates (Bock & Aitkin, 1981) were used] and then computing product-moment correlations. These estimates were then disattenuated to remove the bias caused by the presence of substantial measurement error.

Results

The three models were compared in terms of the usefulness of the information they provided. For the UM and MM joint analyses, they were compared in terms of model-data fit. The log-like lihood statistics (G^2), often called the likelihood ratio statistic, for the MM and the UM were 5,103.05 and 5,137.15, respectively; Akaike's information criterion (AIC; Akaike, 1977) for the UM and MM were 5,184.05 and 5, 183.15, respectively. Thus, the model-data fit of the MM (which estimated 27 parameters) was not a significant improvement over that of the UM (22 parameters) because the AIC were almost identical and the difference in the G^2 (7.1 with 5 degrees of freedom) was not statistically significant. (The CM was not included because it was fit as three independent models with identifying constraints on the item parameters, so that it was not a submodel of the MMs.)

Table 3 shows the estimated population means and variances and a reliability index for each of the latent dimensions. The reliability is the ratio of the observed variance in the EAP estimates to the estimated true variance. The CM and MM produced statistically equivalent estimates of the three latent means and variances (e.g., the mean was $-.29$ for the multistructural dimension for the CM and was $-.30$ for that same dimension for the MM). Note that the reliabilities of the dimensions were much higher for the MM. These higher reliabilities follow from the fact that the EAP method used to estimate abilities in the MM was able to draw on information in all three latent dimensions to improve on the ability predictions for each of the single dimensions. The UM produced only a single mean, variance, and reliability. The reliability was slightly higher than that for the separate subscales in the MM.

Table 4 shows estimates of the correlations between the latent dimensions estimated using the CM and the MM. For the CM, both the attenuated and disattenuated correlation estimates are reported; however, for the MM, only the direct estimates of the latent correlations are reported. The very high correlation estimates are consistent with the adequacy of a UM for these data. These results suggest that, at least for these data, the SOLO levels appeared to measure equivalent dimensions of performance.

Table 3
Estimated Population Means and Variances and Reliabilities for Each of the Latent Dimensions

Model	Dimension	Mean	SE	Variance	SE	Reliability
Unidimensional	Composite	−.27	.08	1.04	.14	.78
Consecutive	Unistructural	.64	.09	.94	.20	.32
	Multistructural	−.29	.10	1.21	.24	.59
	Relational	−1.08	.09	.59	.16	.35
Multidimensional	Unistructural	.64	.09	.95	.21	.73
	Multistructural	−.30	.10	1.02	.25	.76
	Relational	−1.09	.09	.81	.18	.64

Conclusions

The MRCMLM is a mathematically tractable and flexible multidimensional model that produces parameter estimates that are readily interpretable. Further work is necessary to assist test constructors and users who have a need for clear guidance and concrete examples concerning the utilities and limitations of MIRT models. Tools and interpretation strategies also are needed to help in the interpretation of the parameters in multidimensional models. In the case of the model described here, some of this guidance is available in Wang (1994) and Wang, Wilson, & Adams (1996).

Table 4
Correlations Between the Latent Dimensions (UM = Unistructural With Multistructural, UR = Unistructural With Relational, MR = Multistructural With Relational)

Model	UM	UR	MR
Consecutive (Attenuated)	.70	.41	.56
Consecutive (Disattenuated)	1.00	1.00	1.00
Multidimensional	.97	.85	.91

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol III; pp. 143–166). Norwood NJ: Ablex.

Adams, R. J., Wilson, M. R., & Wu, M. L. (in press). Multilevel item response modelling: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics.*

Akaike, H. (1977). On entropy maximisation principle. In P. R. Krischnaiah (Ed.), *Applications of statistics* (pp. 27–41). New York: North Holland.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, 32,* (Series B), 283–301.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3–16.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357–374.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37–48.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 technical report* (Research Rep. No. 15-TR-20). Princeton NJ: Educational Testing Service.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New

York: Academic Press.
Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443–459.
Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16,* 129–147.
Davey, T., & Hirsch, T. M. (1991, May). *Concurrent and consecutive estimates of examinee ability profiles.* Paper presented at the Annual Meeting of the Psychometric Society, New Brunswick NJ.
de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183–196.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39,* (Series B), 1–38.
Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Tests and achievement tests.* New York: College Entrance Examination Board.
Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515.
Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.
Fischer, G. H., & Pononcy, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59,* 177–192.
Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13,* 373–389.
Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* Unpublished doctoral dissertation. University of Twente, The Netherlands.
Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1; pp. 236–258). Norwood NJ: Ablex.
Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635–659.
Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 71–94). Vancouver BC: Educational Research Institute of British Columbia.
Hoijtink, H. (1995). Linear and repeated measures models for the person parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 203–214). New York: Springer-Verlag.
Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149–176.
Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.
Luecht, R. M., & Miller, R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16,* 279–293.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.
Masters, G. N. (1988). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 11–50). Plenum: New York.
Mislevy, R. J., & Bock, R. D. (1983). *BILOG II: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.
Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16,* 237–248.
Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G. H. Fischer & D. Laming, *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 249–263). New York: Springer-Verlag.
Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago, The University of Chicago Press.
Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4,* 321–334.
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.
Volodin, N., & Adams, R. J. (1995). *Identifying and estimating a D-dimensional Rasch model.* Unpublished manuscript, Australian Council for Educational Research, Camberwell, Victoria, Aus-

tralia.

Wang, W. C. (1994). *Implementation and application of the multidimensional random coefficients multinomial logit.* Unpublished doctoral dissertation, University of California, Berkeley.

Wang, W. C., Wilson, M. R., & Adams, R. J. (1996). *Implications and applications of the multidimensional random coefficients multinomial logit model.* Unpublished manuscript, Australian Council for Educational Research, Camberwell, Victoria, Australia.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and non- compensatory two-dimensional data on unidimensional IRT estimation. *Applied Psychological Measurement, 12,* 239–252.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.

Whitely, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

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

Wilson, M. (1988). Detecting breakdowns in local independence using a family of Rasch models. *Applied Psychological Measurement, 12,* 353–364.

Wilson, M. R. (1992). The partial order model: An extension of the partial credit model. *Applied Psychological Measurement, 16,* 309–325.

Wilson, M. R., & Adams, R. J. (1993). Marginal maximum likelihood estimation for the ordered partition model. *Journal of Educational Statistics, 18,* 69–90.

Wilson, M. R., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60,* 181– 198.

Wilson, M. R., & Wang, W. C. (in press). Complex composites: Issues that arise in combining different models of assessment. *Applied Psychological Measurement, 19,* 51–72.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wu, M. L., & Adams, R. J. (1993, April). *Simulating parameter recovery for the random coefficients multinomial logit.* Paper presented at the Sixth International Objective Measurement Workshop, Atlanta.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1995). *MATS: Multi-aspect test software.* Melbourne, Australia: Australian Council for Educational Research.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56,* 589–600.

Acknowledgments

All of the authors were supported by NSF Grant RED-9255272. The first author was partially supported by a National Academy of Education Spencer Fellowship. The third author was supported by a dissertation fellowship from the Chiang Ching-Kuo Foundation.

Author's Address

Send requests for reprints or further information to Ray Adams, Australian Council for Educational Research, Private Bag 55, Camberwell, Victoria, 3124 Australia. Email: adams@acer.edu.au.

Editor's Note

This article is part of the Special Issue on Multidimensional Item Response Theory published as Volume 20 Number 4, December 1996, edited by Terry Ackerman. The article is published in this issue due to space limitations in the December issue.