Research Report

Estimating Multidimensional Item Response Models With Mixed Structure

Jinming Zhang

Research & Development

April 2005
RR-05-04

Estimating Multidimensional Item Response Models With Mixed Structure

Jinming Zhang
ETS, Princeton, NJ

April 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This study derived an expectation-maximization (EM) algorithm for estimating the parameters of multidimensional item response models. A genetic algorithm (GA) was developed to be used in the maximization step in each EM cycle. The focus of the EM-GA algorithm developed in this paper was on multidimensional items with *mixed structure*. Simulated item response data were generated and then estimated by a computer program based on the EM-GA algorithm. The simulation results demonstrate that the EM-GA algorithm is a very promising approach in estimating multidimensional item response model parameters.

Key words: Genetic algorithm, GA, EM-GA algorithm, ASSEST, estimation, multidimensional item response theory, MIRT, mixed structure, approximate simple structure

Acknowledgments

Portions of this article were presented at the annual meeting of the National Council on Measurement in Education, April 2005. The research was supported by ETS. The opinions expressed herein are solely those of the author and do not necessarily represent those of ETS. The author would like to thank Ting Lu for her help in developing the computer program and for her comments and suggestions.

1. Introduction

Educational or psychological tests usually have several target content areas or subscales to measure. For example, the Graduate Record Examinations® (GRE®) General Test measures analytical writing, verbal, and quantitative skills. These tests are typically composed of several sections or subsets of items measuring different subscales. Items measuring the same content area are assumed to be unidimensional; that is, these tests are multidimensional with simple structure. In practice, however, some items may actually be contaminated in the sense that knowledge in other subscales is helpful for an examinee to get correct answers for these items, although the test is designed to have simple structure. Moreover, a test framework may explicitly require some items to measure more than one subscale in its assessment. For example, the National Assessment Governing Board (NAGB; 1994, p. 13) stated in the Mathematics Framework for the 1996 National Assessment of Educational Progress (NAEP) that some items in the assessment "should have major elements drawn from more than one strand." Such a test has no simple structure.

In this paper, an item simply measuring one subscale (content strand or skill) is called a *pure* item, and an item measuring more than one subscale is a *mixed* item. If a test consists of pure items only, it is a simple structure test; otherwise, it is called a *mixed structure test*. In other words, a mixed structure test contains items measuring several subscales, such as a mathematics item that measures both algebra and geometry and items measuring only one subscale as well. The mixed structure assumption is the natural generalization of simple structure and at the same time satisfies the request from some test frameworks that some items should measure several subscales. The content/categorical nature of many test specifications should typically yield mixed structure tests. In practice, some mixed items, according to their contexts, measure one subscale to a greater extent than other subscales. Usually a parameter estimation program can confirm that in its output. However, response data may be very noisy. Thus, sometimes it is preferable to add some constraints on certain mixed items during calibration so that their estimation results are consistent with their contexts. Such mixed items are called *semimixed* in this paper. For example, a one-subscale dominated item may be treated as a semimixed item. In other words, the treatment of

a mixed item as semi-mixed is equivalent to using the prior information that this item mainly measures one subscale. To model item response data from a mixed structure test, multidimensional item response theory (MIRT) should be used.

There is literature on multidimensional item estimation. NOHARM (normal ogive harmonic analysis related method) (Fraser, 1988; Fraser & McDonald, 1988) and TESTFACT (test scoring, item statistics, and item factor analysis) (Wilson, Wood, & Gibbons, 1991) are the two most commonly used multidimensional item response estimation programs. NOHARM uses common factor analysis methodology to estimate item parameters for both unidimensional and multidimensional normal ogive models, while TESTFACT applies full-information factor analysis methodology. These two programs yield similar results (Miller, 1991). With sample sizes over 1,000 and test lengths long enough, these programs have been found to give stable parameter estimates that can be used for a number of applications. However, they are not ready to be applied to the mixed structure case mentioned above because of their limitations. For example, NOHARM can deal with two-parameter multidimensional normal ogive models with mixed structure but cannot deal with three-parameter models that are usually required to model multiple-choice items. Moreover, neither program can deal with semi-mixed items. Therefore, it is necessary to develop appropriate analysis procedures for dealing with the mixed structure case in order to adequately reflect the intent of test frameworks.

The main purpose of this paper is to develop an algorithm to calibrate items of multidimensional tests with mixed structure. In Section 2, multidimensional item response models are introduced and the expectation-maximization (EM) algorithm used to search for marginal maximum likelihood estimates (MMLE) is derived. A genetic algorithm (GA) is developed in Section 3. The GA is used in the maximization step in each EM cycle. A GA is a computational algorithm that incorporates ideas from genetics and/or evolution (e.g., breeding, mutation, crossover, and survival of the fittest) to solve optimization problems. Section 4 presents some simulation results using the EM-GA algorithm developed in Sections 2 and 3, and Section 5 provides further discussion.

2. MIRT Models and the EM Algorithm

Suppose there is a test with n dichotomously scored items, and X_i is the score on item i for a randomly selected examinee from a certain population. The *item response function* (IRF) is defined as the probability of answering an item correctly for a randomly selected examinee with ability vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, where d is the number of dimensions of the test. That is, $P_i(\boldsymbol{\theta}) = P(X_i = 1 \mid \boldsymbol{\theta})$.

One widely used MIRT model is the multidimensional compensatory three-parameter logistic (M3PL) model. Its IRF is

$$P_i(\boldsymbol{\theta}) = c_i + (1 - c_i)\frac{1}{1 + \exp\{-1.7(\sum_{k=1}^{d} a_{ik}\theta_k - d_i)\}} \tag{1}$$

where

a_{ik} $(k = 1, \ldots, d)$ are the discrimination parameters (nonnegative and not all zero),

d_i is the parameter that is related to the difficulty of item i, and

c_i is the lower-asymptote parameter $(0 \leq c_i < 1)$.

All discrimination parameters are required to be nonnegative so that the IRF is a nondecreasing function of all abilities. When c_i is set to be zero, the M3PL model becomes a multidimensional two-parameter logistic (M2PL) model (see Reckase, 1985; Reckase & McKinley, 1991). The M3PL model (1) is often reparametrized as

$$P_i(\boldsymbol{\theta}) = c_i + (1 - c_i)\frac{1}{1 + \exp\{-1.7\sum_{k=1}^{d} a_{ik}(\theta_k - b_i)\}} \tag{2}$$

where $b_i = d_i / \sum_{k=1}^{d} a_{ik}$, so that the difficulty parameter is directly comparable with that in the usual expression of a unidimensional three-parameter logistical (3PL) model (see Lord, 1980).

There is a discrimination parameter for each dimension being modeled, but only one parameter relating to the overall item difficulty in the model (1) or (2). Using a distinct difficulty parameter for each separate dimension would lead to an indeterminate (i.e., unidentifiable) solution, hence that is statistically inappropriate. Since the term in the exponent is a linear combination of abilities, high ability values on some dimensions can

be compensated for low ability values on the other dimensions. For item response data modeled by compensatory models, the discrimination parameter vector is the unique factor to determine the dimensional structure of an item. When there is only one nonzero discrimination parameter, the M3PL model (2) becomes a unidimensional 3PL model.

Theoretically, any coordinate system can be used in MIRT. However, the constraint that all discrimination parameters are nonnegative requires that a coordinate system should be chosen such that all discrimination vectors lie in the first quadrant, as shown in Figure 1. Figure 1 graphically represents a two-dimensional mixed structure test, where θ_1^* and θ_2^* are canonical coordinate axes in the sense that they are not correlated. In practice, many test frameworks often stipulate that their test items measure several subscales (content strands, or content areas). These target subscales, shown as θ_1 and θ_2, are preferable in use as the coordinate axes, rather than θ_1^* and θ_2^*, since such a coordinate system has substantive meaning, as would be the case with algebra and geometry in a mathematics test. These subscales are anchored by items selected by test developers and/or through dimensionality analysis. Note that θ_1 and θ_2 are positively correlated. The anchor items lying along the axes are the pure items and other items are mixed items, as discussed in Section 1. Note that if a canonical coordinate system, such as θ_1^* and θ_2^* shown in Figure 1, is used, then every item in Figure 1 measures both θ_1^* and θ_2^* and thus, is a mixed item with respect to θ_1^* and θ_2^*. When this paper discusses pure or mixed items (i.e., items measuring one subscale or more than one subscale), it always refers to the target subscales.

In theory, items determine what subscales a test measures. Therefore, θ_1 and θ_2 could just be two composites of latent variables. In Figure 1, for example, θ_1 and θ_2 are expected to be in alignment with two sets of pure items. The substantive meanings of θ_1 and θ_2 should be determined by the pure items. If the first set of pure items measures algebra, then θ_1 is the algebra subscale. That is the reason why pure items are sometimes called *anchor* items; they anchor the subscales. When item i is pure, either a_{i1} or a_{i2} is zero in (2) for $d = 2$. Under the simple structure assumption, there is one and only one nonzero discrimination parameter for each item. In other words, each item in a simple structure test measures θ_1 or θ_2 only. If a_{i1} is relatively large (say $a_{i1} = 1.0$) and a_{i2} is relatively small

(say $a_{i2} = 0.5$), the ith item mainly measures θ_1 since the probability variation due to a change in θ_1 value is larger than that due to a change in θ_2. According to their contexts, some (mixed) items measure one subscale to a greater extent than the other. If such prior information is used during item parameter estimation, these items are considered to be semimixed, as explained in Section 1. That is, when a mixed item is specified as semimixed, a constraint, $a_1 > a_2$ or $a_1 < a_2$, is imposed in the model during item parameter estimation, depending on whether the item is a first subscale predominant item or a second subscale predominant item.

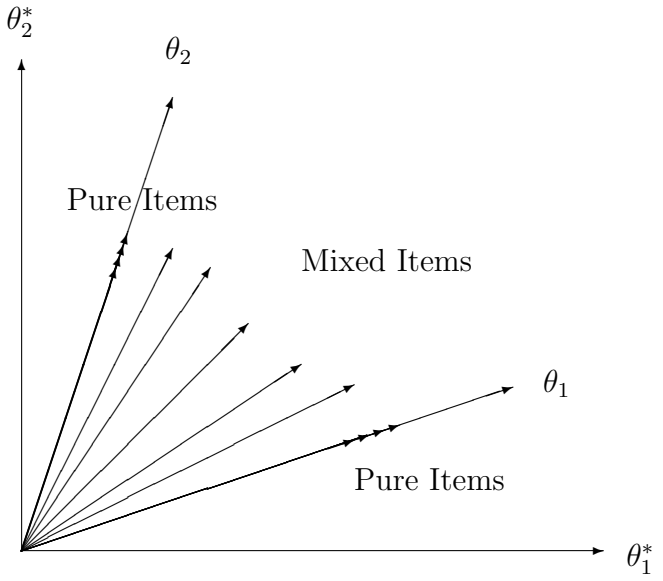

Figure 1. **A two-dimensional test with mixed structure.**

The marginal maximum likelihood estimation approach (Bock & Aitkin, 1981) is used in this paper. The prior distribution of the abilities is assumed to be a multivariate normal distribution. Without loss of generality, one can standardize the abilities so that they have means of zero and variances of one. The correlation coefficients between the abilities are unknown parameters that also need to be estimated. Generally speaking, there is a linear indeterminacy of the ability scales in the M2PL/M3PL models; that is, any nonsingular linear transformation can be made for abilities. In this paper, the target subscales are used as coordinate axes, which is determined by pure items, and hence no linear transformation

is allowed here. In practice, every item measuring only one content area should be regarded as a pure item, unless there is some evidence it is not, based on either its context or the results of dimensionality analysis. Note that under the simple structure assumption, every item is a pure item.

The marginal likelihood function can be calculated as below. By local independence, the joint probability of a particular response pattern $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})$ across a set of n items given the jth examinee's $\boldsymbol{\theta}_j$ is

$$P(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Gamma}) = \prod_{i=1}^{n} P_i(\boldsymbol{\theta}_j)^{x_{ij}}(1 - P_i(\boldsymbol{\theta}_j))^{1-x_{ij}}, \quad j = 1, 2, \ldots, m, \tag{3}$$

where $P_i(\boldsymbol{\theta})$ is the IRF, $\boldsymbol{\Gamma}$ is the set of item parameters, and m is the number of examinees. For a randomly sampled examinee j from a population with the prior distribution $\varphi(\boldsymbol{\theta} \mid \boldsymbol{\rho})$, the marginal probability of an observed response pattern $\boldsymbol{x}_j$ is

$$P(\boldsymbol{x}_j \mid \boldsymbol{\rho}, \boldsymbol{\Gamma}) = \int \cdots \int P(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Gamma}) \varphi(\boldsymbol{\theta}_j \mid \boldsymbol{\rho}) d\boldsymbol{\theta}_j,$$

where $\boldsymbol{\rho}$ denotes the (unknown) correlation coefficients between subscales. The marginal likelihood function of the response patterns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ from m randomly sampled examinees is given by

$$L(\boldsymbol{\rho}, \boldsymbol{\Gamma}; \mathbf{X}) = \prod_{j=1}^{m} P(\boldsymbol{x}_j \mid \boldsymbol{\rho}, \boldsymbol{\Gamma}),$$

where

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix}$$

is the response data matrix. The natural logarithm of the marginal likelihood function is

$$\log L(\boldsymbol{\rho}, \boldsymbol{\Gamma}; \mathbf{X}) = \sum_{j=1}^{m} \log \int \cdots \int P(\boldsymbol{x}_j \mid \boldsymbol{\theta}, \boldsymbol{\Gamma}) \varphi(\boldsymbol{\theta} \mid \boldsymbol{\rho}) d\boldsymbol{\theta}. \tag{4}$$

For convenience, the j subscript on $\boldsymbol{\theta}$ is dropped in the above equation and in subsequent expressions whenever $\boldsymbol{\theta}$ is a dummy variable.

The MMLEs of the unknown parameters $\boldsymbol{\rho}$ and $\boldsymbol{\Gamma}$ are based on the information provided by the response data $\mathbf{X}$ and are obtained by maximizing the log marginal likelihood function (4). However, it is not feasible to compute the MMLE directly by maximizing (4) unless the number of items is small (e.g., 6). For example, there are $2n$ unknown parameters in (4), even in the simplest case that the test is unidimensional and all items are modeled by two-parameter logistic functions. The MMLE method becomes practical only when an EM algorithm is used (see Bock & Aitkin, 1981). The EM algorithm is an iterative method for finding maximum likelihood estimates of parameters for probability models (Dempster, Laird & Rubin, 1977). Each iteration consists of two steps: the E step (expectation step) and the M step (maximization step). In the E step, the conditional expectation (i.e., the posterior expectation) of the log likelihood function for complete data is computed given provisional estimated item parameters and the observed (incomplete) data. In the M step, the posterior expectation is maximized with respect to parameters to obtain updated estimated parameters. This process is repeated until a certain convergence criterion is met (e.g., likelihood function value and all item parameter estimates become stable to some degree). For details, see Baker (1992) and Tanner (1996).

In IRT settings, latent traits are treated as latent (missing) data in the EM algorithm. The observed data $\mathbf{X}$ and latent traits are regarded as the complete data. Thus, the log likelihood function of complete data is

$$\begin{aligned}
\log f(\mathbf{X}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \mid \boldsymbol{\rho}, \boldsymbol{\Gamma}) &= \log \prod_{j=1}^{m} P(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Gamma}) \varphi(\boldsymbol{\theta}_j \mid \boldsymbol{\rho}) \\
&= \sum_{j=1}^{m} \log P(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Gamma}) + \sum_{j=1}^{m} \log \varphi(\boldsymbol{\theta}_j \mid \boldsymbol{\rho}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} [x_{ij} \log P_i(\boldsymbol{\theta}_j) + (1 - x_{ij}) \log(1 - P_i(\boldsymbol{\theta}_j))] \\
&\quad + \sum_{j=1}^{m} \log \varphi(\boldsymbol{\theta}_j \mid \boldsymbol{\rho}). \qquad (5)
\end{aligned}$$

The posterior ability distribution given $\boldsymbol{x}_j$ is

$$p(\boldsymbol{\theta}_j \mid \boldsymbol{x}_j, \boldsymbol{\rho}, \boldsymbol{\Gamma}) = \frac{P(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Gamma}) \varphi(\boldsymbol{\theta}_j \mid \boldsymbol{\rho})}{\int \cdots \int P(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\Gamma}) \varphi(\boldsymbol{\theta}_j \mid \boldsymbol{\rho}) d\boldsymbol{\theta}_j}. \qquad (6)$$

By (5) and (6), the posterior expectation of the log likelihood function for complete data, given provisional estimated parameters (i.e., $\boldsymbol{\rho}^*$ and $\boldsymbol{\Gamma}^*$) and the observed data $\mathbf{X}$, can be obtained:

$$\begin{aligned} Q(\boldsymbol{\Gamma}, \boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*) &= E[\log f(\mathbf{X}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \mid \boldsymbol{\rho}, \boldsymbol{\Gamma}) \mid \mathbf{X}, \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*] \\ &= \sum_{i=1}^{n} \left[\int \cdots \int \sum_{j=1}^{m} x_{ij} p(\boldsymbol{\theta} \mid \boldsymbol{x}_j, \boldsymbol{\rho}^*, \boldsymbol{\Gamma}^*) \log P_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \right. \\ &\qquad \left. + \int \cdots \int \sum_{j=1}^{m} (1 - x_{ij}) p(\boldsymbol{\theta} \mid \boldsymbol{x}_j, \boldsymbol{\rho}^*, \boldsymbol{\Gamma}^*) \log(1 - P_i(\boldsymbol{\theta})) d\boldsymbol{\theta} \right] \\ &\quad + \int \cdots \int \sum_{j=1}^{m} p(\boldsymbol{\theta} \mid \boldsymbol{x}_j, \boldsymbol{\rho}^*, \boldsymbol{\Gamma}^*) \log \varphi(\boldsymbol{\theta} \mid \boldsymbol{\rho}) d\boldsymbol{\theta} \\ &= \sum_{i=1}^{n} Q_i(\boldsymbol{\gamma}_i; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*) + Q_0(\boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*) \end{aligned} \tag{7}$$

where $\boldsymbol{\gamma}_i$ is the set of item parameters of item i (in $P_i(\boldsymbol{\theta})$),

$$\begin{aligned} Q_i(\boldsymbol{\gamma}_i; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*) &= \int \cdots \int R_i(\boldsymbol{\theta}) \log P_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\quad + \int \cdots \int [R_0(\boldsymbol{\theta}) - R_i(\boldsymbol{\theta})] \log(1 - P_i(\boldsymbol{\theta})) d\boldsymbol{\theta}, \end{aligned} \tag{8}$$

$$Q_0(\boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*) = \int \cdots \int R_0(\boldsymbol{\theta}) \log \varphi(\boldsymbol{\theta} \mid \boldsymbol{\rho}) d\boldsymbol{\theta}, \tag{9}$$

$$R_i(\boldsymbol{\theta}) = \sum_{j=1}^{m} x_{ij} p(\boldsymbol{\theta} \mid \boldsymbol{x}_j, \boldsymbol{\rho}^*, \boldsymbol{\Gamma}^*), \tag{10}$$

$$R_0(\boldsymbol{\theta}) = \sum_{j=1}^{m} p(\boldsymbol{\theta} \mid \boldsymbol{x}_j, \boldsymbol{\rho}^*, \boldsymbol{\Gamma}^*). \tag{11}$$

The objective of the M step in an EM cycle is to find the $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\rho}}$ such that $Q(\boldsymbol{\Gamma}, \boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ is maximized at $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\rho}}$. According to (7), the maximization of $Q(\boldsymbol{\Gamma}, \boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ can be carried out via the maximization of both $Q_0(\boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ and $Q_i(\boldsymbol{\gamma}_i; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ for each item singly. That is, one only need to find a $\widehat{\boldsymbol{\rho}}$ such that $Q_0(\boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ is maximized at $\widehat{\boldsymbol{\rho}}$, and a $\widehat{\boldsymbol{\gamma}}_i$ such that $Q_i(\boldsymbol{\gamma}_i; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ is maximized at $\widehat{\boldsymbol{\gamma}}_i$ for $i = 1, 2, \ldots, n$.

The EM algorithm developed in this paper is briefly described as follows:

1. Initialization. The initial values for parameters are either provided by the user or

generated by the program as default values. The default initial values for parameters are set as follows:

- The correlation coefficients between abilities are set either to 0.5 or at the value of the sample correlations between observed raw (number-correct) scores of respective subtests.
- The discrimination parameters are set at 1, except for secondary dimensions in semimixed cases. The discrimination parameters for secondary dimensions of semimixed items are set to be 0.5 (e.g., for a two-dimensional first-scale dominated semimixed item, the default values of a_1 and a_2 are set to be 1 and 0.5, respectively).
- The guessing parameters for 3PL models are set at 0.2.
- The difficulty parameters are set at the values of the inverse of the logistic function at the mean scores of the respective items.

Note that the program can be used sequentially; that is, the estimated item parameters from the previous run (as a whole or as a part) can be used as the initial item parameters for the next run.

2. The E step.

   Compute $R_i(\boldsymbol{\theta})$ and $R_0(\boldsymbol{\theta})$ in (10) and (11), given provisional item parameter estimates $\boldsymbol{\Gamma}^*$ and provisional correlation estimates $\boldsymbol{\rho}^*$.

3. The M step.

   - The $Q_i(\boldsymbol{\gamma}_i; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ function in (8) is maximized with respect to $\boldsymbol{\gamma}_i$ to obtain the updated estimate for $\boldsymbol{\gamma}_i$, given provisional item parameter estimates $\boldsymbol{\Gamma}^*$ and provisional correlation estimates $\boldsymbol{\rho}^*$ for $i = 1, 2, \ldots, n$.
   - The $Q_0(\boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ function in (9) is maximized with respect to $\boldsymbol{\rho}$ to obtain the updated estimate for ρ given provisional item parameter estimates $\boldsymbol{\Gamma}^*$ and provisional correlation estimates $\boldsymbol{\rho}^*$.

4. The stopping rule.

   If the log marginal likelihood is unchanged from the previous cycle, the parameter estimation process has converged and the process is terminated. Otherwise, steps 2 and 3 are repeated.

The Newton-Raphson method is used to maximize $Q_0(\boldsymbol{\rho}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ with respect to $\boldsymbol{\rho}$. For easy presentation, this paper uses a two-dimensional case as an example to describe the method. In a two-dimensional case, the prior distribution of latent traits is

$$\varphi(\boldsymbol{\theta} \mid \rho) \equiv \varphi(\theta_1, \theta_2 \mid \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\theta_1^2 - 2\rho\theta_1\theta_2 + \theta_2^2\right) \right\}, \tag{12}$$

where ρ is the correlation coefficient between the two latent traits. Some useful derivatives are as follows:

$$\begin{aligned}
\frac{\partial}{\partial\rho}\left(\log\varphi(\boldsymbol{\theta} \mid \rho) \right) &= \frac{\rho}{1-\rho^2} + \frac{1}{(1-\rho^2)^2}[(1+\rho^2)\theta_1\theta_2 - \rho(\theta_1^2+\theta_2^2)], \\
\frac{\partial^2}{\partial\rho^2}\left(\log\varphi(\boldsymbol{\theta} \mid \rho) \right) &= \frac{1+\rho^2}{(1-\rho^2)^2} + \frac{1}{(1-\rho^2)^2}[2\rho\theta_1\theta_2 - (\theta_1^2+\theta_2^2)], \\
&\quad + \frac{4\rho}{(1-\rho^2)^3}[(1+\rho^2)\theta_1\theta_2 - \rho(\theta_1^2+\theta_2^2)].
\end{aligned}$$

Therefore,

$$\begin{aligned}
Q_0'(\rho; \boldsymbol{\Gamma}^*, \rho^*) &= \int\int R_0(\boldsymbol{\theta}) \, \frac{\partial}{\partial\rho}\left(\log\varphi(\boldsymbol{\theta} \mid \rho) \right) d\boldsymbol{\theta} \\
&= \frac{m\rho}{1-\rho^2} + \frac{1}{(1-\rho^2)^2}[(1+\rho^2)A - \rho B], \\
Q_0''(\rho; \boldsymbol{\Gamma}^*, \rho^*) &= \int\int R_0(\boldsymbol{\theta}) \, \frac{\partial^2}{\partial\rho^2}\left(\log\varphi(\boldsymbol{\theta} \mid \rho) \right) d\boldsymbol{\theta} \\
&= \frac{m(1+\rho^2)}{(1-\rho^2)^2} + \frac{1}{(1-\rho^2)^2}(2\rho A - B) + \frac{4\rho}{(1-\rho^2)^3}[(1+\rho^2)A - \rho B],
\end{aligned}$$

where

$$\begin{aligned}
A &= \int\int \theta_1\theta_2 R_0(\theta_1, \theta_2) d\theta_1 d\theta_2, \\
B &= \int\int (\theta_1^2 + \theta_2^2) R_0(\theta_1, \theta_2) d\theta_1 d\theta_2.
\end{aligned}$$

The iterative formula for the Newton-Raphson method is

$$\rho_{k+1} = \rho_k - \frac{Q_0'(\rho_k; \mathbf{\Gamma}^*, \rho^*)}{Q_0''(\rho_k; \mathbf{\Gamma}^*, \rho^*)}, \quad k = 1, 2, \ldots \tag{13}$$

The Newton-Raphson method could also be used in the maximization step to find the $\boldsymbol{\gamma}_i$ such that $Q_i(\boldsymbol{\gamma}_i; \mathbf{\Gamma}^*, \boldsymbol{\rho}^*)$ is maximized. However, there are several item parameters with many constraints under the mixed structure assumption. In such a case, the Newton-Raphson method is not effective and feasible. Besides, in general, the log likelihood function may have multiple modes and/or saddle points (see Tanner, 1996). In order to improve upon the EM algorithm, a GA is developed and used in the maximization step. That is, a GA is used to find $\widehat{\boldsymbol{\gamma}}$ such that $Q_i(\boldsymbol{\gamma}; \mathbf{\Gamma}^*, \boldsymbol{\rho}^*)$ is maximized at $\widehat{\boldsymbol{\gamma}}$, given $\mathbf{\Gamma}^*$ and $\boldsymbol{\rho}^*$.

3. Genetic Algorithm in the Maximization Step of the EM Algorithm

The basic notion of a GA is to mimic the principles of natural evolution. A GA starts with a set (called *population*) of potential solutions (each solution is called an *individual*) to the problem at hand. Then, it stochastically optimally selects individuals as parents of the next generation and lets the selected individuals clone, mutate, and combine some of their components to form new individuals. This process proceeds over successive generations until one cannot find an individual better than the optimal individual one has gotten so far. Currently, GAs have been quite successfully applied to many optimization problems (see Michalewicsz, 1994; Jiang & Tang, 1998; Zhang & Stout, 1999).

A GA is developed in this paper to find $\widehat{\boldsymbol{\gamma}}$ such that $Q_i(\boldsymbol{\gamma}; \mathbf{\Gamma}^*, \boldsymbol{\rho}^*)$ is maximized at $\widehat{\boldsymbol{\gamma}}$, given $\mathbf{\Gamma}^*$ and $\boldsymbol{\rho}^*$. For concreteness, a two-dimensional 3PL model is used to describe the algorithm. Recall that the IRF of a two-dimensional item is

$$P(\theta_1, \theta_2) = c + (1 - c)\frac{1}{1 + \exp\{-1.7[a_1(\theta_1 - b) + a_2(\theta_2 - b)]\}}, \tag{14}$$

where a_1 and a_2 are nonnegative and not all zero. Note that the subscript of item sequence is not used here for convenience since this section focuses on one item. Thus, the notation in this section is different from other sections in this paper. There are five types of items with different dimensional structure that can be represented by model (14). They are

first-subscale or second-subscale pure items, first-subscale or second-subscale dominated semimixed items, and mixed items. When an item is a first-subscale (or second-subscale) pure item, a_2 (or a_1) is fixed to be zero in (14), which becomes a unidimensional 3PL model. When an item is specified as a first-subscale (or second-subscale) dominated item, a constraint, $a_1 > a_2 \geq 0$ (or $a_2 > a_1 \geq 0$), will be imposed. The algorithm treats each of those five types of items differently by imposing different constraints or no constraint at all. Below a first-subscale dominated item is used as an example to describe the algorithm since the treatments for the other types of items are either similar or less complicated.

In the GA, any vector $\boldsymbol{\gamma} = (a_1, a_2, b, c)$ with $a_1 > a_2 \geq 0$ is regarded as a potential solution that may maximize $Q_i(\boldsymbol{\gamma}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ for a first-subscale dominated item. The population in this GA consists of those vectors (i.e., individuals). The population size, denoted as M, is the number of individuals in the population. Recall that in a GA a potential solution is called an individual and a specified whole set of individuals is called the population. Generally speaking, the larger the population size, the greater the chance of getting the optimal solution. However, there is a tradeoff between accuracy and efficiency. When the population size is larger, the computer program will take a longer time to run. Typically, the population size can be chosen from any number between 30 to 100 (the default number is 80), and once chosen, it is fixed during the evolution process.

There are two basic genetic operators in the GA: mutation and crossover. Generally, mutation arbitrarily alters some genes of a selected individual. For details, see Michalewicsz (1994). A mutation operator used in this paper is the following transformation: $\boldsymbol{\gamma} = (a_1, a_2, b, c)$ mutates to be $\boldsymbol{\gamma}^* = (a_1^*, a_2^*, b^*, c^*)$ with

$$\begin{aligned} a_1^* &= a_1 + \varepsilon_1, \\ a_2^* &= a_2 + \varepsilon_2, \\ b^* &= b + \varepsilon_3, \\ c^* &= c + \varepsilon_4, \end{aligned}$$

where ε_i are independent random variables. All of the ε_i are chosen to be normal with means of zero and are truncated with certain lower and upper bounds so that the new

parameters after mutation are within reasonable ranges. After each mutation, the program will check whether $\boldsymbol{\gamma}^*$ is a legitimate individual or not (e.g., check whether it satisfies constraints, such as $a_1^* > a_2^* \geq 0$). Some treatments will be applied to the new individual if it does not satisfy the constraints. For example, if the constraint, $a_1^* > a_2^* \geq 0$, does not hold, a positive random number will be subtracted from a_2^* so that the new a_2^* is between zero and a_1^*. Note that this is the only place where all constraints are enforced. Then, the original individual is replaced by the new one.

In this paper, a crossover operator is defined to be a binary transformation. Suppose that $\boldsymbol{\gamma}_1 = (a_{11}, a_{12}, b_1, c_1)$ and $\boldsymbol{\gamma}_2 = (a_{21}, a_{22}, b_2, c_2)$ are two individuals. The crossover operator randomly exchanges some components between $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ to form two new individuals. For example, one possible result after a crossover is two new individuals $\boldsymbol{\gamma}_1^* = (a_{11}, a_{12}, b_2, c_1)$ and $\boldsymbol{\gamma}_2^* = (a_{21}, a_{22}, b_1, c_2)$, for which the third components b_1 and b_2 have been exchanged with each other. Then, the original pair is discarded.

The GA developed in this paper is described as follows:

Step 1. Initialization.

First, the initial potential solution $\boldsymbol{\gamma}_0$ is the provisional estimate of the item parameter vector from the previous M step or the initial item parameter vector if the GA is being used in the first cycle of the EM algorithm. See Section 2 for details about how to provide the initial values for item parameters in the EM algorithm. Then, let $\boldsymbol{\gamma}_0$ mutate repeatedly and independently to generate some individuals (about 85% of the population size). The initial population consists of these individuals and some clones of $\boldsymbol{\gamma}_0$ (individuals exactly the same as $\boldsymbol{\gamma}_0$). Finally, $Q_i(\boldsymbol{\gamma}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ is calculated for each individual in the initial population. The Q_i value is the measure of the fitness of the individual: The larger the Q_i value is, the more suitable (better) the individual is.

Step 2. Selection.

The selection process is based on fitness of individuals. First, all individuals in the current population are ordered according to their Q_i values from the smallest to the

largest. Then, M individuals are sampled randomly *with replacement* from the population (M is the population size). The probability of each individual being selected is proportional to its rank. The better the individual, the greater the chance of being selected. The M selected individuals constitute an updated population, which will be used in the next step. Note that there is a large chance that good individuals are selected more than once while some bad individuals may never be selected.

Step 3. Breeding.

The crossover and mutation operators are used to produce new individuals. For each individual there is some chance of being selected to mutate (e.g., 50% chance) and of being crossed over with another selected individual (e.g., 30% chance). The probabilities of being able to mutate and being crossed are determined empirically. Usually they are set to be between 30% to 70% for mutation and between 10% to 40% for crossover. Once set, they are fixed throughout the process for every individual. See Michalewicsz (1994) for more discussion about the selection of those probabilities. Those new individuals, together with the remaining individuals (i.e., their clones), constitute the next generation population.

Step 4. Evaluation.

Calculate $Q_i(\boldsymbol{\gamma}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ for each individual in the population from Step 3. Record the maximum $Q_i(\boldsymbol{\gamma}; \boldsymbol{\Gamma}^*, \boldsymbol{\rho}^*)$ value. If this value does not change (increase) much when compared to the previous generations' maximum values, then the GA stops. Otherwise Steps 2, 3, and 4 are repeated.

The idea behind a GA is to do what nature does. What the selection and breeding steps in the GA do is to let good individuals have a relatively large chance to produce offspring, while some bad individuals produce no offspring. In other words, at each generation, relatively good individuals have a relatively large chance of being reproduced, while relatively bad individuals may die without reproduction. However, the GA still allows some lucky bad individuals to be selected and lets them survive and reproduce. After all, a bad individual itself, or with another individual, might produce very good offspring.

Therefore, the GA performs a *multidirectional* search to find the global optimal solution. After several generations, the GA converges and the best individual in the last generation is claimed to be a near optimal solution.

4. Simulation Studies

The EM-GA algorithm presented in Sections 2 and 3 has been implemented in a Fortran program called ASSEST, short for Approximate Simple Structure ESTimation. To investigate its performance, ASSEST was used to estimate item parameters with simulated unidimensional and two-dimensional response data. Item parameter estimates obtained from ASSEST were compared with the true ones. The root mean-squared error (RMSE) of the estimated parameters is commonly used as a criterion for the recovery of item parameters in simulation studies. The RMSE is the square root of the average of the squared deviations of estimated parameters from the corresponding true ones.

In practice, the estimates of item parameters are usually treated as fixed in any further analysis of response data, such as estimating abilities of examinees. In the process of such an analysis, the IRF is more directly relevant than item parameters themselves since most operational statistical analysis is based on the likelihood function formed by the IRFs. In addition, different sets of item parameters may produce very close item characteristic curves or surfaces. Therefore, it is more appropriate to check the closeness of estimated IRF (curves or surfaces) to the true IRF than the item parameter estimates to the true values. The estimated IRF, denoted as $\hat{P}_i(\boldsymbol{\theta})$, is the IRF with the estimated item parameters. The RMSE of $\hat{P}_i(\boldsymbol{\theta})$ is the Euclidean distance between the estimated IRF and its corresponding true IRF:

$$D_i = \sqrt{E\{[\hat{P}_i(\boldsymbol{\Theta}) - P_i(\boldsymbol{\Theta})]^2\}},$$

where the expectation E is respect to the latent ability vector $\boldsymbol{\Theta}$. Or

$$D_i = \sqrt{\int [\hat{P}_i(\boldsymbol{\theta}) - P_i(\boldsymbol{\theta})]^2 \varphi(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}) d\boldsymbol{\theta}}, \tag{15}$$

where $\varphi(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$ is the density function of the latent ability vector and $\boldsymbol{\Sigma}$ is its correlation matrix. Clearly, the smaller the RMSE, the better the estimator is. If the density function

of the latent ability vector is $\varphi(\theta_1, \theta_2 \mid \rho)$ given by (12) for a two-dimensional case, then the RMSE of an estimated IRF is

$$D_i = \sqrt{\int\int [\hat{P}_i(\theta_1, \theta_2) - P_i(\theta_1, \theta_2)]^2 \varphi(\theta_1, \theta_2 \mid \rho) d\theta_1 d\theta_2}. \tag{16}$$

For a unidimensional case, (15) can be simplified as

$$D_i = \sqrt{\int [\hat{P}_i(\theta) - P_i(\theta)]^2 \varphi(\theta) d\theta}, \tag{17}$$

where $P_i(\theta)$ is the unidimensional 2PL or 3PL model and $\varphi(\theta)$ is the standard normal distribution in this paper.

4.1 A Unidimensional Case

Although it is designed to estimate item parameters for multidimensional models, the EM-GA algorithm can also be used to estimate item parameters for unidimensional models. Here, a single simulated case is presented as an example to show the performance of ASSEST in a unidimensional case.

In this study, the estimated item parameters of 25 dichotomously scored physical science items from the 1996 NAEP advanced science assessment were used as true item parameters to generate simulated response data. Among those 25 items, there are 7 short constructed response items modeled by 2PL models in the NAEP operational analysis and 18 multiple-choice items modeled by 3PL models. These item parameters were obtained from the NAEP BILOG/PARSCALE program, which is an item parameter estimation program that combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs. For convenience, BILOG was used to represent the NAEP BILOG/PARSCALE program in this paper.

The number of simulated examinees was 2,000. Examinees' ability scores were generated independently from a standard normal distribution. Simulated response data were generated using the following (standard) IRT method. Given ability score θ_j (or $\boldsymbol{\theta}_j = (\theta_{1j}, \theta_{2j})$ for a two-dimensional case below), first calculate the probability of answering item i correctly by examinee j, $p_{ij} = P_i(\theta_j)$, using the true item parameters from Table 1

(or Table 2 for two-dimensional cases). Then independently generate a random number r from the $(0, 1)$ uniform distribution. If $r < p_{ij}$, then a correct response was obtained for examinee j on item i; otherwise, an incorrect response was obtained.

Both ASSEST and BILOG were used to recover the item parameters with a simulated 2000×25 response data set. In both runs, the ability distribution was fixed to be standard normal. The two sets of item parameter estimates from ASSEST and BILOG, along with the true item parameters, are given in Table 1. The last two columns present the RMSEs of estimated IRFs from both ASSEST and BILOG using (17). Their means and standard deviations are about the same, 0.017 and 0.009, respectively (see the last two numbers in the last two columns of Table 1). The second to last row presents the sample correlation coefficients between the estimated and true item parameters, which range from 0.8256 to 0.9911, indicating they are highly correlated. The last row of Table 1 gives the RMSE of estimated item parameters within the test for the discrimination, difficulty, and lower-asymptote parameters; that is, $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\gamma}_i - \gamma_i)^2}$ for $\gamma_i = a_i$, b_i or c_i. Note that for items modeled by 2PL models, the lower-asymptote parameters were fixed at zero. Thus, here $n = 25$ for the a and b parameters and $n = 18$ for the c parameter.

Figure 2 shows four typical examples of item characteristic curves (ICC). There are three curves in each plot for each selected item: the solid curve is the true ICC, the dotted one is the estimated ICC from ASSEST, and the dashed one is the estimated ICC from BILOG. The first plot (Item 6) gives a typical good-fit example for both ASSEST and BILOG. The second plot (Item 7) presents an example of good-fit for BILOG but not for ASSEST, while the third plot (Item 15) shows that ASSEST provides a better estimation than BILOG. And the fourth plot (Item 8) presents an example of bad-fit from both programs.

Table 1

True Estimated Item Parameters and RMSE of Estimated IRFs From ASSEST and BILOG for a Unidimensional Test

	a			b			c			RMSE of IRF	
Item	T	A	B	T	A	B	T	A	B	A	B
1	0.310	0.426	0.333	0.450	0.727	0.184	0.268	0.342	0.235	0.014	0.013
2	0.696	0.729	0.685	-0.014	-0.092	-0.174	0.292	0.271	0.259	0.011	0.014
3	0.828	0.724	0.922	-0.797	-1.037	-0.675	0.195	0.080	0.202	0.010	0.021
4	0.528	0.544	0.619	1.673	1.621	1.573	0.000	0.000	0.000	0.004	0.020
5	0.442	0.416	0.503	0.265	0.305	0.314	0.000	0.000	0.000	0.009	0.022
6	0.918	0.988	0.921	-0.027	-0.047	0.037	0.000	0.000	0.000	0.013	0.019
7	1.159	1.634	1.212	1.215	1.204	1.238	0.107	0.127	0.112	0.030	0.007
8	1.248	1.613	0.798	2.132	2.093	2.318	0.199	0.216	0.196	0.014	0.028
9	1.071	1.398	0.867	1.243	1.300	1.248	0.281	0.314	0.252	0.020	0.015
10	0.777	1.266	0.599	1.513	1.557	1.440	0.289	0.318	0.237	0.035	0.016
11	0.556	0.599	0.625	-2.060	-1.900	-1.843	0.241	0.298	0.206	0.005	0.020
12	0.774	0.848	0.784	-0.855	-0.900	-0.846	0.000	0.000	0.000	0.022	0.002
13	0.586	0.626	0.577	-0.334	-0.190	-0.573	0.281	0.351	0.217	0.014	0.014
14	0.694	0.818	0.555	0.160	0.159	0.193	0.211	0.197	0.201	0.026	0.026
15	0.740	0.768	0.886	0.569	0.627	0.493	0.366	0.380	0.313	0.004	0.044
16	0.894	0.670	0.863	0.924	0.695	1.086	0.200	0.120	0.221	0.028	0.022
17	0.566	0.475	0.503	1.182	1.327	1.211	0.000	0.000	0.000	0.022	0.018
18	0.759	1.071	0.549	1.816	1.847	2.078	0.271	0.297	0.228	0.022	0.027
19	0.863	0.993	0.868	1.878	1.763	1.905	0.000	0.000	0.000	0.015	0.005
20	1.404	1.537	1.202	2.067	2.061	2.248	0.112	0.125	0.116	0.010	0.017
21	0.837	1.126	0.883	1.487	1.448	1.498	0.112	0.131	0.115	0.024	0.007
22	0.855	0.715	0.829	-1.648	-1.984	-1.741	0.250	0.176	0.203	0.016	0.002
23	0.498	0.509	0.515	-0.847	-0.457	-0.927	0.241	0.391	0.201	0.023	0.008
24	0.603	0.609	0.501	0.276	0.545	-0.176	0.352	0.409	0.209	0.017	0.019
25	0.423	0.415	0.482	0.077	0.079	0.075	0.000	0.000	0.000	0.002	0.018
Mean	0.762	0.861	0.724	0.494	0.510	0.488	0.238	0.253	0.207	0.017	0.017
SD	0.266	0.380	0.223	1.155	1.159	1.195	0.075	0.107	0.051	0.009	0.009
Corr[a]	-	0.900	0.866	-	0.990	0.991	-	0.826	0.890	-	-
RMSE	-	0.205	0.135	-	0.159	0.159	-	0.062	0.047	-	-

Note. T is true. A is ASSEST. B is BILOG.

[a] Sample correlation between estimated and true item parameters.

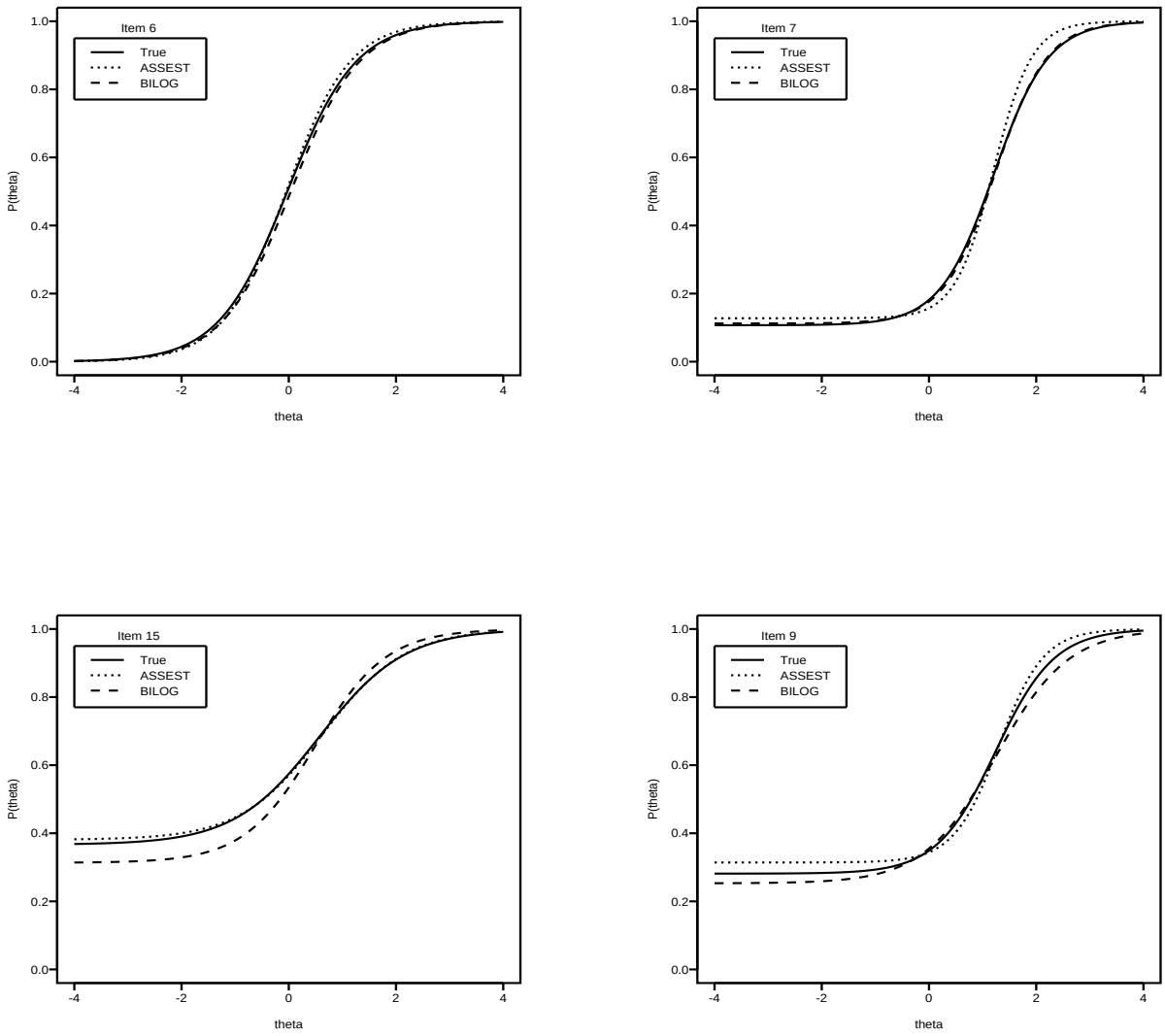

Figure 2. **Four typical examples of estimated ICCs from ASSEST and BILOG.**

Although ASSEST seems to overestimate item parameters whereas BILOG seems to underestimate them for the simulated response data according to Table 1, these two programs gave approximately the same level of accuracy in estimating item parameters and IRFs. Under the unidimensional assumption, ASSEST is similar to the program developed by Jiang and Tang (1998). Interested readers can find more unidimensional results in that paper.

4.2 Two-Dimensional Cases

The numbers of items in the following simulated two-dimensional tests are 30, 46, or 62. The estimated item parameters of dichotomous items from the analysis of the 1998 NAEP grade 4 reading assessment (see Appendix E of Allen, Donoghue, & Schoeps, 2001) were used as true item parameters in this simulation. There are 31 dichotomous items measuring the first subscale of reading for literary experience and 32 dichotomous items measuring the second subscale of reading to gain information. A bad item in the second scale with $b = 3.921$ was dropped from our simulation studies. Therefore, there is a total of 62 items with 35 multiple-choice and 27 constructed-response items. Note that all NAEP items are modeled by unidimensional models and the whole test has simple structure. To make a test with mixed structure, 22 items were selected to become mixed items by adding positive values as their other discrimination parameters. Thus, the total number of items is 62 with 40 pure and 22 mixed items. For completeness, these item parameters are presented in Table 2. For tests with 30 or 46 items, the first 15 or 23 items from each subscale were chosen. For instance, the items in the 30-item test are Items 1-15 and Items 32-46 in Table 2. Note that a shorter test is a subtest of a longer test. The numbers of mixed items in the 30-item and 46-item tests are 10 and 16, respectively. Note also that no semimixed items were specified and used in the simulation study.

The number of simulated examinees was 1,000, 3,000, or 5,000 in this study. Examinees' true ability scores were generated independently from a bivariate normal distribution with mean of 0, variance of 1, and correlation of 0.5 or 0.8. Note that the estimated correlation coefficients between subscales in NAEP assessments are usually around 0.8 (see Allen, Carlson, & Zelenak, 1999; Allen et al., 2001), and the typical correlation coefficient is 0.5 between math and verbal in an achievement test with math and verbal sections, such as the SAT®.

In summary, the following three factors were considered in this simulation study:

1. the number of items: 30, 46, or 62 with 10, 16, or 22 mixed items, respectively;
2. the number of simulated examinees: 1,000, 3,000, or 5,000; and
3. the correlation coefficient between two subscales: 0.5 or 0.8.

Given these three factors, there were 18 combinations or cases in this simulation. For each case, ASSEST was applied to a simulated response data set to get parameter estimates. This process was repeated 100 times for each case.

Table 2

Item Parameters Used in the Two-Dimensional Cases

First subscale					Second subscale				
Item	a_1	a_2	b	c	Item	a_1	a_2	b	c
1	0.623	0.000	-0.872	0.000	32	0.000	0.269	-0.904	0.000
2	1.506	0.000	-0.495	0.215	33	0.000	0.941	0.401	0.264
3	0.920	0.000	1.008	0.000	34	0.000	0.793	0.642	0.247
4	0.607	0.000	0.712	0.251	35	0.000	1.032	0.507	0.248
5	1.052	0.000	1.009	0.000	36	0.000	1.172	0.645	0.000
6	1.288	0.000	0.554	0.190	37	0.000	0.533	-0.835	0.218
7	1.798	0.000	-0.899	0.248	38	0.000	0.877	-0.523	0.000
8	0.754	0.000	0.015	0.000	39	0.000	1.203	0.257	0.165
9	1.342	0.000	-0.457	0.175	40	0.000	0.761	-1.242	0.000
10	0.763	0.000	-0.284	0.000	41	0.000	1.104	-0.155	0.247
11	1.110	0.565	0.148	0.244	42	0.412	0.619	-1.113	0.000
12	1.025	0.734	0.107	0.000	43	0.657	1.154	0.645	0.000
13	1.228	0.487	0.259	0.247	44	0.406	1.464	0.774	0.138
14	0.647	0.550	-1.008	0.000	45	0.578	1.536	1.192	0.000
15	0.520	0.387	-1.425	0.000	46	0.373	0.597	1.341	0.000
16	0.951	0.000	-0.864	0.319	47	0.000	2.300	0.416	0.264
17	0.757	0.000	-0.630	0.000	48	0.000	0.562	-0.073	0.237
18	0.832	0.000	1.118	0.000	49	0.000	0.970	0.906	0.000
19	1.472	0.000	1.204	0.167	50	0.000	0.883	-1.015	0.310
20	1.859	0.000	0.213	0.265	51	0.000	1.261	1.084	0.206
21	1.123	0.821	1.057	0.000	52	0.503	0.597	-0.206	0.156
22	1.133	0.528	0.916	0.297	53	0.438	0.938	-1.691	0.294
23	1.374	0.442	0.307	0.269	54	0.631	1.086	-0.060	0.000
24	0.504	0.000	-0.932	0.247	55	0.000	0.795	-0.238	0.000
25	1.415	0.000	0.891	0.271	56	0.000	1.414	-0.608	0.275
26	2.303	0.000	0.609	0.418	57	0.000	0.838	-0.076	0.000
27	0.814	0.000	0.306	0.000	58	0.000	1.185	-0.590	0.312
28	0.966	0.000	-1.318	0.244	59	0.000	1.031	-0.310	0.000
29	0.506	0.476	-1.272	0.000	60	0.395	0.579	-0.688	0.276
30	1.029	0.368	0.327	0.300	61	0.783	0.970	-0.502	0.270
31	0.721	0.533	-1.193	0.247	62	0.569	1.002	-0.530	0.000

In this simulation, the RMSE of estimated parameters was calculated in a regular way, which is different from the RMSE used in the unidimensional case without replications. Let γ_i represent a parameter of item i, a_{ik}, b_i, or c_i, and $\widehat{\gamma}_{ij}$ be the estimate of γ_i from the jth replications for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Here n is the number of items and J is the number of replications ($J = 100$ in the simulation study). For each item parameter, define

$$\text{RMSE}(\gamma_i) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\widehat{\gamma}_{ij} - \gamma_i)^2}.$$

These RMSEs are usually further summarized as the average of the RMSEs, ARMSE, for each type of item parameter. There are four types of item parameters, two discrimination parameters a_k ($k = 1, 2$), the difficulty parameter b, and the lower-asymptote parameter c for multiple-choice items. The ARMSE is defined as

$$\text{ARMSE}(\gamma) = \frac{1}{\#S_\gamma}\sum_{i\in S_\gamma}\text{RMSR}(\gamma_i) = \frac{1}{\#S_\gamma}\sum_{i\in S_\gamma}\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\widehat{\gamma}_{ij} - \gamma_i)^2}, \tag{18}$$

where γ represents one of the four types of item parameters, S_γ is the set of item sequence numbers where parameter γ needs to be estimated and $\#S_\gamma$ is the number of elements in S_γ. Note that only those item parameters that need to be estimated are included in the ARMSE calculation; that is, item parameters that are fixed as zero are excluded in the calculation. If γ is the lower-asymptote parameter, for example, then $S_c = \{i\text{: item } i \text{ is a multiple-choice item, } 1 \le i \le n\}$ and $\#S_c$ is the number of items modeled by M3PL models.

Similarly, the average of RMSEs of estimated IRFs among the items in a test across all replications will be used as an overall measure of the accuracy of the estimation. The overall average $\bar{d} = \frac{1}{nJ}\sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij}$ is called the ARMSE of estimated IRFs, where d_{ij} is the RMSE of estimated ith IRF in the jth replication given by (16).

Table 3 presents the ARMSEs of the estimated item parameters and IRFs. Columns 4-7 are the ARMSEs for the four types of item parameters and the last column is the ARMSE of the estimated IRFs. As expected, when the correlation between subscales and the number of items are fixed, the ARMSE decreases as the number of examinees increases. That is, the larger the number of examinees, the better the estimates are. When both the numbers of examinees and items are fixed, the ARMSE of IRFs increases as the correlation

between subscales increases from 0.5 to 0.8. This is also true for the ARMSEs of the estimated item parameters in most cases considered here.

Table 3

ARMSE of Estimated Item Parameters and IRFs Based on 100 Replications

n	m	ρ	a_1	a_2	b	c	IRF
	1,000	0.5	0.1536	0.1486	0.1242	0.0705	0.0244
		0.8	0.1764	0.1687	0.1240	0.0709	0.0248
30	3,000	0.5	0.0856	0.0828	0.0772	0.0422	0.0149
		0.8	0.1059	0.0969	0.0791	0.0435	0.0153
	5,000	0.5	0.0698	0.0636	0.0644	0.0358	0.0119
		0.8	0.0846	0.0810	0.0639	0.0350	0.0123
	1,000	0.5	0.1578	0.1483	0.1278	0.0710	0.0246
		0.8	0.1834	0.1725	0.1285	0.0699	0.0253
46	3,000	0.5	0.0922	0.0892	0.0870	0.0465	0.0159
		0.8	0.1111	0.1030	0.0877	0.0448	0.0166
	5,000	0.5	0.0744	0.0710	0.0728	0.0371	0.0134
		0.8	0.0904	0.0820	0.0758	0.0366	0.0142
	1,000	0.5	0.1630	0.1431	0.1306	0.0757	0.0244
		0.8	0.1905	0.1742	0.1311	0.0738	0.0254
62	3,000	0.5	0.0967	0.0895	0.0893	0.0494	0.0157
		0.8	0.1180	0.1075	0.0906	0.0467	0.0167
	5,000	0.5	0.0782	0.0721	0.0746	0.0401	0.0128
		0.8	0.0959	0.0883	0.0784	0.0387	0.0141

Note. n is the number of items. m is the number of examinees. ρ is the correlation coefficient between two subscales.

As in the unidimensional case, the sample correlation coefficients between the estimated and true item parameters were calculated for the discrimination, difficulty, and

lower-asymptote parameters within each replication. Their averages, over 100 replications, are presented in Table 4. Note that item parameters fixed to be zero are excluded in the calculation of the sample correlation coefficients. These sample correlation coefficients are pretty high for the discrimination and difficulty parameters, but relatively low for the lower-asymptote parameter. As the number of examinees increases, the sample correlation coefficients between the estimated and true item parameters also increase. However, the number of items in the range considered here has little impact on the average of the sample correlation coefficients between the estimated and true item parameters.

The correlation coefficients between subscales were estimated during the estimation of item parameters. These estimates are pretty close to their corresponding true correlations. The RMSE of estimated correlations based on 100 replications is listed in the last column of Table 5. As shown in Table 5, the largest RMSE is 0.0240, which appears in the 30-item 1,000-examinee 0.5-correlation case, while the smallest RMSE is 0.0073 in the case of 46 items, 5,000 examinees, and 0.8 correlation. Generally speaking, the greater the number of examinees, the better the estimated correlation is. The RMSEs at correlation of 0.8 are smaller than those at correlation of 0.5, which is possibly due the nature of correlation coefficients that have larger variations in the middle than near the upper or lower bounds. However, the impact of test length on the estimation of the correlation is not so straightforward. When both the number of examinees and the correlation between two subscales are fixed, 46-item tests have the smaller RMSE than 30-item and 62-item tests five out of six cases.

It should be noted that all results reported above were obtained under the assumption that the test structure was known; that is, both pure and mixed items were correctly identified. Recall that a pure item has one and only one positive discrimination parameter. When an item is specified as a pure item in an ASSEST run, one discrimination parameter is estimated while the other discrimination parameter is fixed at zero in a two-dimensional case. When an item is specified as a mixed item, both discrimination parameters are estimated. In practice, it is essential that a dimensionality analysis be performed before the calibration. However, it is still quite possible that some items will be classified incorrectly. To explore the consequence of the misclassification, all the above ASSEST runs were replicated using the same response data sets but with some items misspecified. Two

situations are investigated in this paper: (a) all mixed items are specified as pure items in their dominant subscale or (b) some pure items are specified as mixed items.

Table 4

The Average of Sample Correlations Between Estimated and True Item Parameters Based on 100 Replications

n	m	ρ	a_1	a_2	b	c
30	1,000	0.5	0.9320	0.9203	0.9816	0.4159
		0.8	0.9172	0.8923	0.9828	0.4225
	3,000	0.5	0.9773	0.9721	0.9927	0.6292
		0.8	0.9647	0.9589	0.9918	0.6048
	5,000	0.5	0.9838	0.9838	0.9946	0.6766
		0.8	0.9780	0.9707	0.9945	0.6795
46	1,000	0.5	0.9337	0.9387	0.9827	0.4370
		0.8	0.9161	0.9183	0.9830	0.4601
	3,000	0.5	0.9748	0.9771	0.9918	0.6314
		0.8	0.9660	0.9695	0.9918	0.6527
	5,000	0.5	0.9844	0.9861	0.9947	0.7189
		0.8	0.9783	0.9816	0.9943	0.7286
62	1,000	0.5	0.9335	0.9406	0.9800	0.4802
		0.8	0.9180	0.9156	0.9800	0.5019
	3,000	0.5	0.9765	0.9780	0.9900	0.6533
		0.8	0.9691	0.9695	0.9906	0.6870
	5,000	0.5	0.9851	0.9864	0.9931	0.7270
		0.8	0.9812	0.9810	0.9932	0.7477

Note. n is the number of items. m is the number of examinees. ρ is the correlation coefficient between two subscales.

Table 5

Average of True and Estimated Correlations Between Abilities and RMSE of Estimated Correlations Based on 100 Replications

n	m	ρ	True abilities' Corr.	(their SD)	Average est. Corr.	(their SD)	RMSE of Est. corr.
30	1,000	0.5	0.5019	(0.0222)	0.5032	(0.0342)	0.0240
		0.8	0.8000	(0.0101)	0.7876	(0.0192)	0.0221
	3,000	0.5	0.5008	(0.0139)	0.5017	(0.0214)	0.0146
		0.8	0.7991	(0.0063)	0.7890	(0.0114)	0.0146
	5,000	0.5	0.4999	(0.0112)	0.4994	(0.0175)	0.0133
		0.8	0.7996	(0.0060)	0.7906	(0.0098)	0.0122
46	1,000	0.5	0.5019	(0.0222)	0.5098	(0.0318)	0.0210
		0.8	0.8000	(0.0101)	0.8028	(0.0167)	0.0149
	3,000	0.5	0.5008	(0.0139)	0.5100	(0.0201)	0.0156
		0.8	0.7991	(0.0063)	0.8029	(0.0100)	0.0091
	5,000	0.5	0.4999	(0.0112)	0.5077	(0.0156)	0.0132
		0.8	0.7996	(0.0060)	0.8038	(0.0081)	0.0073
62	1,000	0.5	0.5019	(0.0222)	0.5187	(0.0294)	0.0236
		0.8	0.8000	(0.0101)	0.8118	(0.0142)	0.0166
	3,000	0.5	0.5008	(0.0139)	0.5181	(0.0181)	0.0199
		0.8	0.7991	(0.0063)	0.8121	(0.0085)	0.0145
	5,000	0.5	0.4999	(0.0112)	0.5144	(0.0146)	0.0172
		0.8	0.7996	(0.0060)	0.8121	(0.0078)	0.0135

Note. n is the number of items. m is the number of examinees. ρ is the correlation coefficient between two subscales.

First, ASSEST was applied to the same simulated response data as the original simulation study but with all mixed items treated (incorrectly) as pure items. That is, these simulated data were calibrated again as simple structure tests. The new results were

summarized in Table 6. The true abilities' correlations and their standard deviations in column 4 of Table 6 are the same as those in Table 5 because the same response data were used. Those numbers are placed in Table 6 for the readers' convenience. The new ARMSE of the estimated IRFs have increased (comparing the last column of Table 6 to that of Table 3). For instance, the ARMSE of the estimated IRFs is now 0.0415 in the case of 30 items, 5,000 examinees and 0.5 correlation, whereas the corresponding one in Table 3 is 0.0119, where all mixed items were correctly identified. In most cases considered in the simulation study, the new ARMSE of the estimated IRFs has more than doubled in size over those from the original runs.

When mixed items were incorrectly specified as pure items, the estimated correlations between subscales had a big positive bias. The new average estimated correlation between subscales was much larger than the original average of estimated correlations and the true correlation (see the fifth columns in Tables 5 and 6). For example, the new average estimated correlation is 0.7411 in the case of 30 items, 5,000 examinees and 0.5 correlation, while the original average estimated one is 0.4994 and the true one is 0.4999. Because the mixed items were incorrectly specified as pure items, ASSEST actually calibrated two composites, which were the combinations of the target subscales. These two composite subscales leaned closer to each other than the target subscales did. Not surprisingly, the correlation between the two subscales was overestimated. This issue was further investigated. For details, see Zhang (2004). Consequently, the RMSE of the estimated correlations also dramatically increased (comparing the second to last column of Table 6 with the last column of Table 5). For instance, the RMSE of the estimated correlations is now 0.2415 in the case of 30 items, 5,000 examinees, and 0.5 correlation, while the original one with the correct identification of mixed items is 0.0133. These results indicate that it is inappropriate to treat a test with mixed structure as a simple structure test. However, the consequence of wrong mixed item identifications is less severe for a highly correlated case ($\rho = 0.8$) than a moderately correlated case ($\rho = 0.5$).

Table 6

Average of True and Estimated Correlations Between Abilities and RMSE of Estimated Correlations and IRFs With All Mixed Items as Pure Based on 100 Replications

n	m	ρ	True abilities' Corr.	(their SD)	Average est. Corr.	(their SD)	RMSE of Est. corr.	ARMSE of Est. IRF
30	1,000	0.5	0.5019	(0.0222)	0.7442	(0.0223)	0.2431	0.0493
		0.8	0.8000	(0.0101)	0.8959	(0.0103)	0.0966	0.0385
	3,000	0.5	0.5008	(0.0139)	0.7435	(0.0121)	0.2428	0.0433
		0.8	0.7991	(0.0063)	0.8970	(0.0065)	0.0982	0.0323
	5,000	0.5	0.4999	(0.0112)	0.7411	(0.0103)	0.2415	0.0415
		0.8	0.7996	(0.0060)	0.8972	(0.0051)	0.0978	0.0303
46	1,000	0.5	0.5019	(0.0222)	0.7617	(0.0192)	0.2604	0.0496
		0.8	0.8000	(0.0101)	0.9109	(0.0082)	0.1113	0.0391
	3,000	0.5	0.5008	(0.0139)	0.7617	(0.0112)	0.2610	0.0439
		0.8	0.7991	(0.0063)	0.9112	(0.0051)	0.1122	0.0329
	5,000	0.5	0.4999	(0.0112)	0.7602	(0.0089)	0.2605	0.0422
		0.8	0.7996	(0.0060)	0.9115	(0.0043)	0.1120	0.0310
62	1,000	0.5	0.5019	(0.0222)	0.7651	(0.0173)	0.2636	0.0492
		0.8	0.8000	(0.0101)	0.9159	(0.0074)	0.1162	0.0387
	3,000	0.5	0.5008	(0.0139)	0.7650	(0.0101)	0.2643	0.0431
		0.8	0.7991	(0.0063)	0.9160	(0.0046)	0.1171	0.0324
	5,000	0.5	0.4999	(0.0112)	0.7626	(0.0084)	0.2629	0.0413
		0.8	0.7996	(0.0060)	0.9161	(0.0038)	0.1166	0.0305

Note. n is the number of items. m is the number of examinees. ρ is the correlation coefficient between two subscales.

Next, consider the opposite situation to that above: some pure items are specified as mixed items. Specifically, 8 pure items, Items 9, 10, 20, 28, 40, 41, 51, and 59, were specified as mixed items in the next simulation study, whenever they were included in the

tests. Therefore, there were 4, 6, and 8 pure items that were specified as mixed items in the 30-item, 46-item, and 62-item tests, respectively. The results from the third simulation study are summarized in Tables 7 through 9. These tables are parallel to Tables 3-5, where all items were specified correctly according to their structure. Those pure items that were specified as mixed items got slightly worse estimation results than the results from the original simulation. The ARMSEs of the estimated IRFs all slightly increased when some pure items were specified as mixed items (see the last columns of Tables 3 and 7). For the pure items that were specified as mixed, the estimates of the discrimination parameter that would have been zero had the items been appropriately treated as pure turned out to be very small (typically around 0.25 or less). The correlations between true and estimated item parameters have no big changes (compare Table 8 with Table 4). These correlations even increased for discrimination parameters from the original simulation mainly due to the fact that the estimates of the zero discrimination parameters are pretty close to zero. The RMSE of estimated correlations between subscales only increased for 30-item cases but decreased for 62-item cases (see the last columns of Tables 5 and 9). Overall, all results remain approximately the same when comparing Tables 7-9 with Tables 3-5, which indicates the robustness of ASSEST. Note that in this simulation study the percentage of misspecified pure items is 20%. If more pure items are misspecified, then the estimation results will become worse.

Note that the purpose of the EM-GA algorithm is to search for MMLEs of item parameters. Thus, the ideal criterion to judging the algorithm is to check whether the maximum marginal likelihood value occurs at the estimates found by ASSEST, or whether these estimates are really optimal points of the marginal likelihood function. The dilemma is that the EM-GA algorithm is actually used because it is extremely difficult to search for these optimal points. Instead, one may compare the marginal likelihood value found by ASSEST with the corresponding marginal likelihood value at the true parameters. If the marginal likelihood value at parameter estimates is larger than the corresponding marginal likelihood value at the true parameters when models are correctly specified, then the search algorithm used to obtain the parameter estimates may be judged good enough.

Table 7

ARMSE of Estimated Item Parameters and IRFs With Some Pure Items As Mixed Based on 100 Replications

n	m	ρ	a_1	a_2	b	c	IRF
	1,000	0.5	0.1506	0.1423	0.1265	0.0718	0.0255
		0.8	0.1775	0.1710	0.1251	0.0711	0.0263
30	3,000	0.5	0.0850	0.0817	0.0822	0.0438	0.0163
		0.8	0.1136	0.1035	0.0818	0.0446	0.0169
	5,000	0.5	0.0684	0.0625	0.0680	0.0356	0.0132
		0.8	0.0905	0.0866	0.0682	0.0367	0.0140
	1,000	0.5	0.1552	0.1433	0.1308	0.0720	0.0258
		0.8	0.1826	0.1725	0.1316	0.0704	0.0267
46	3,000	0.5	0.0902	0.0881	0.0910	0.0466	0.0173
		0.8	0.1122	0.1068	0.0914	0.0451	0.0180
	5,000	0.5	0.0721	0.0695	0.0773	0.0377	0.0147
		0.8	0.0907	0.0862	0.0791	0.0373	0.0153
	1,000	0.5	0.1575	0.1388	0.1330	0.0755	0.0254
		0.8	0.1856	0.1717	0.1334	0.0734	0.0265
62	3,000	0.5	0.0932	0.0876	0.0932	0.0493	0.0168
		0.8	0.1137	0.1068	0.0944	0.0477	0.0178
	5,000	0.5	0.0754	0.0703	0.0791	0.0411	0.0139
		0.8	0.0918	0.0874	0.0810	0.0391	0.0149

Note. n is the number of items. m is the number of examinees. ρ is the correlation coefficient between two subscales.

Table 8

The Average of Sample Correlations Between Estimated and True Item Parameters With Some Pure Items as Mixed Based on 100 Replications

n	m	ρ	a_1	a_2	b	c
	1,000	0.5	0.9471	0.9421	0.9816	0.4062
		0.8	0.9299	0.9141	0.9827	0.4280
30	3,000	0.5	0.9826	0.9799	0.9922	0.6076
		0.8	0.9686	0.9661	0.9917	0.5911
	5,000	0.5	0.9886	0.9883	0.9947	0.6818
		0.8	0.9802	0.9766	0.9943	0.6596
	1,000	0.5	0.9467	0.9530	0.9826	0.4288
		0.8	0.9276	0.9324	0.9826	0.4505
46	3,000	0.5	0.9812	0.9827	0.9919	0.6248
		0.8	0.9713	0.9739	0.9917	0.6512
	5,000	0.5	0.9882	0.9895	0.9945	0.7067
		0.8	0.9816	0.9835	0.9941	0.7220
	1,000	0.5	0.9486	0.9554	0.9800	0.4812
		0.8	0.9332	0.9326	0.9800	0.5078
62	3,000	0.5	0.9826	0.9837	0.9899	0.6566
		0.8	0.9758	0.9759	0.9902	0.6747
	5,000	0.5	0.9888	0.9900	0.9928	0.7178
		0.8	0.9852	0.9848	0.9931	0.7430

Note. n is the number of items. m is the number of examinees. ρ is the correlation coefficient between two subscales.

Table 9

Average of True and Estimated Correlations Between Abilities and RMSE of Estimated Correlations With Some Pure Items As Mixed Based on 100 Replications

n	m	ρ	True abilities' Corr.	(their SD)	Average est. Corr.	(their SD)	RMSE of Est. corr.
	1,000	0.5	0.5019	(0.0222)	0.4844	(0.0362)	0.0322
		0.8	0.8000	(0.0101)	0.7698	(0.0220)	0.0369
30	3,000	0.5	0.5008	(0.0139)	0.4873	(0.0226)	0.0210
		0.8	0.7991	(0.0063)	0.7727	(0.0125)	0.0286
	5,000	0.5	0.4999	(0.0112)	0.4873	(0.0183)	0.0189
		0.8	0.7996	(0.0060)	0.7755	(0.0105)	0.0256
	1,000	0.5	0.5019	(0.0222)	0.4945	(0.0330)	0.0225
		0.8	0.8000	(0.0101)	0.7888	(0.0183)	0.0194
46	3,000	0.5	0.5008	(0.0139)	0.5003	(0.0205)	0.0130
		0.8	0.7991	(0.0063)	0.7918	(0.0110)	0.0116
	5,000	0.5	0.4999	(0.0112)	0.5000	(0.0157)	0.0110
		0.8	0.7996	(0.0060)	0.7941	(0.0087)	0.0083
	1,000	0.5	0.5019	(0.0222)	0.5063	(0.0308)	0.0185
		0.8	0.8000	(0.0101)	0.8008	(0.0157)	0.0128
62	3,000	0.5	0.5008	(0.0139)	0.5107	(0.0180)	0.0141
		0.8	0.7991	(0.0063)	0.8046	(0.0091)	0.0089
	5,000	0.5	0.4999	(0.0112)	0.5085	(0.0145)	0.0126
		0.8	0.7996	(0.0060)	0.8059	(0.0081)	0.0083

Note. n is the number of items. m is the number of examinees. ρ is the correlation coefficient between two subscales.

The marginal likelihood values at the true parameters were calculated in all three two-dimensional simulation studies. There are 18 different cases considered in each simulation. Thus, there are 1,800 ASSEST runs in each simulation. In all of the 3,600 ASSEST runs where all mixed items were correctly identified (the first and third simulation studies), the marginal likelihood values at the estimated item parameters found by ASSEST are larger than the corresponding marginal likelihood values at the true parameters. This result indicates that ASSEST has achieved the limitation of the marginal maximum likelihood approach if items are correctly identified or are not very badly specified (e.g., only 20% of the pure items were specified as mixed items). At the same time, it also shows that the EM-GA algorithm developed in this paper works fine. When some mixed items were incorrectly specified as pure items (i.e., the second simulation), the marginal maximum likelihood values found by ASSEST are all smaller than the corresponding marginal likelihood values at the true parameters for the cases of $\rho = 0.5$ or the number of examinees is 3,000 or 5,000. There are 174 runs out of 1,800 ASSEST runs where the marginal maximum likelihood values found by ASSEST are larger than the corresponding marginal likelihood values at the true parameters; all these happened when $\rho = 0.8$ and the number of examinees is 1,000. This indicates that when the number of examinees is small (e.g., 1,000), the marginal likelihood function is very noisy.

5. Discussion

In this paper, an EM-GA algorithm has been developed to estimate parameters for MIRT models, especially for compensatory models with mixed structure. Simulation studies show that the EM-GA based program, ASSEST, yields quite satisfactory results. Zhang and Lu (2001) compared ASSEST with NOHARM (Fraser & McDonald, 1988) using simulated two-dimensional response data. Their results demonstrate that both ASSEST and NOHARM yield satisfactory estimates of item parameters for compensatory models when the numbers of items and examinees are large enough, and the performance of ASSEST is at least as good as NOHARM for multidimensional compensatory two-parameter logistic models. Recently, Zhang and Stone (2004) reached a similar conclusion when comparing ASSEST with NOHARM. While NOHARM only deals with multidimensional compensatory two-parameter normal ogive models (which are very close to M2PL models

in nature, and their corresponding parameters are comparable), ASSEST not only lets a user choose a M3PL or a M2PL model for each item in a test but also lets a user specify special mixed items, called semimixed, as an option. Recall that a semimixed item, as defined in Section 2, is a one-subscale dominated item; that is, it measures one subscale more heavily than other subscales. Zhang and Lu (2002) further expanded the capacity of the algorithm to estimate the parameters of multidimensional noncompensatory models, introduced by Sympson (1987). Their simulation results showed that ASSEST also yielded quite satisfactory estimation results for the noncompensatory models. The simulation results reported in this paper, as well as other simulated data analyses, demonstrate that the EM-GA algorithm is a very promising approach to estimating MIRT models. The major disadvantage of the EM-GA algorithm is that it requires extensive computational time. The CPU time on a Pentium 2.26 GHz PC for the case of 5,000 examinees and 62 items was about 11 to 15 minutes in the simulation studies. The applications of the EM-GA algorithm to other MIRT models are under investigation.

The main advantage of the mixed structure approach, when compared to an exploratory multidimensional approach, is that the calibrated subscales have substantive meanings, such as algebra and geometry in a mathematics test. The interpretation of the subscales, of course, depends on pure items; subscales are what these pure items measure. Therefore, it is crucial to classify items correctly into pure and mixed items when using the mixed-structure approach to calibrate a test. When an item is specified as a pure or a semimixed item, ASSEST actually uses the prior information on the item, either from its context or from data analysis. A test framework usually specifies what subscales the test tries to measure. Thus, the classification can be typically done by using expert opinion. Expert opinion may also be incorporated with information from a dimensionality analysis of the test item response data. In general, dimensionality analysis together with expert opinion should lead to a satisfactory classification of pure and mixed items used in the mixed-structure approach. In addition, ASSEST can be used in sequence to get an optimal classification of items as mixed or pure. In the first ASSEST run, every item can be considered as a pure item, unless there is some evidence it is not, either from statistical analysis or expert opinion. After the first run, any items specified as mixed items with only one moderate or large discrimination parameter and very small other discrimination parameters in a

compensatory model may be considered as pure items, and any items specified as pure items with bad model-fit may be specified as mixed items in the second run, and so on. In this process, estimated correlations between subscales may serve as an indicator of goodness of classification of pure and mixed items. If some mixed items are incorrectly classified as pure items in calibration, the correlation coefficients actually estimated are larger than the corresponding target correlation coefficients between subscales, according to Theorem 3 of Zhang (2004), which is also confirmed by the second two-dimensional simulation study presented in Section 4. Only when all mixed items are correctly identified are the correlation coefficients to be estimated the target ones. The classification with relatively small estimated correlations among all plausible different mixed-item selections can be regarded as the optimal classification of items. In the process, item contexts and/or contents should always be considered when determining an item to be a pure, a semimixed, or a mixed item.

References

Allen, N., Carlson, J. E., & Zelenak, C. (1999). *The NAEP 1996 technical report* (NCES 1999-452). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Allen, N., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of the EM algorithm. *Psychometrika, 46*, 443-459.

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, Series B, 39,* 1-38.

Fraser, C. (1988). NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent traits theory [Computer software]. Armidale, New South Wales, Australia: The University of New England.

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

Jiang, H., & Tang, L. (1998, April). *New method of calibrating IRT models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Johnson, E. G., & Carlson, J. E. (1994). *The NAEP 1992 technical report* (NCES 94-490). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Michalewicsz, Z. (1994). *Genetic algorithms + data structures = genetic programs.* Berlin: Springer-Verlag.

Miller, T. R. (1991). *Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM estimation program* (Research Rep.

No. ONR 91-2). Iowa City, IA: American College Testing.

Mislevy, R., & Bock, R. D] (1982). BILOG: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software, Inc.

Muraki, E., & Bock, R. D] (1991). PARSCALE: Parameter scaling of rating data [Computer software]. Chicago, IL: Scientific Software, Inc.

National Assessment Governing Board. (1994). *Mathematics Framework for the 1996 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

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.

Sympson, J. B. (1987). A model for testing multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98). Minneapolis: University of Minnesota.

Tanner, M. A. (1996). *Tools for statistical inference.* New York: Springer-Verlag.

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

Zhang, B., & Stone, C. (2004, April). *Direct and indirect estimation of three-parameter compensatory multidimensional item response models.* Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Zhang, J. (2004). Comparison of unidimensional and multidimensional approaches to IRT parameter estimation (ETS RR-04-44). Princeton, NJ: ETS.

Zhang, J., & Lu, T. (2001, April). Evaluating the performance of ASSEST: A new item parameter estimation program for multidimensional item response theory models. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Zhang, J., & Lu, T. (2002, April). *Estimation of multidimensional noncompensatory models.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.