PSYCHOMETRIKA—VOL. 61, NO. 2, 331–354
JUNE 1996

MULTIDIMENSIONAL ADAPTIVE TESTING

DANIEL O. SEGALL

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

Maximum likelihood and Bayesian procedures for item selection and scoring of multidimensional adaptive tests are presented. A demonstration using simulated response data illustrates that multidimensional adaptive testing (MAT) can provide equal or higher reliabilities with about one-third fewer items than are required by one-dimensional adaptive testing (OAT). Furthermore, holding test-length constant across the MAT and OAT approaches, substantial improvements in reliability can be obtained from multidimensional assessment. A number of issues relating to the operational use of multidimensional adaptive testing are discussed.

Key words: computerized adaptive testing, mathematical models, multidimensional adaptive testing, multidimensional item response theory, reliability, tests.

Introduction

Over the past decade, Computerized Adaptive Testing (CAT) has achieved great popularity. Adaptive tests possess several benefits over conventional paper-and-pencil tests. These include increased measurement precision, reduced testing time, standardized instructions, and flexible scheduling of examinees. Most adaptive tests use item selection and scoring algorithms based on Item Response Theory (IRT). To date, these techniques used in operational adaptive testing rely on the assumption of unidimensionality.

A number of investigators have examined the issue of dimensionality in IRT. For the most part, the focus of this work has been on studying the consequences of using unidimensional IRT models in the presence of multidimensional data (Ackerman, 1989, 1991; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Folk & Green, 1989; Harrison, 1986; Reckase, 1979; Reckase, Ackerman, & Carlson, 1988; Way, Ansley, & Forsyth, 1988; Yen, 1984). Two notable exceptions involve the development and evaluation of multidimensional adaptive testing (MAT) estimation procedures by Bloxom and Vale (1987), and Tam (1992). The procedure developed by Bloxom and Vale is a multivariate extension of Owen's sequential Bayesian adaptive updating algorithm (Owen, 1975). The work by Tam evaluated this procedure and five others. Tam compared the estimators using several criteria including precision, test information, and computation time. These studies (Bloxom & Vale, 1987; Tam, 1992) focused on the performance of ability estimation methods—item selection procedures used in these evaluations assumed ideal item pools.

Several multidimensional item selection strategies have been compared using simulated adaptive testing data (Miller, Reckase, Spray, Luecht, & Davey, in press). Examinees and their item responses were simulated according to a two-dimensional latent space, assuming a bivariate normal distribution with uncorrelated abilities. Pro-

The work reported in this paper was sponsored by the Office of Naval Research. The author wishes to thank the three anonymous reviewers for their useful comments on an earlier version of this manuscript. The opinions expressed in this article are those of the Author, are not official and do not necessarily reflect the views of the Navy Department.

Requests for reprints should be sent to Daniel O. Segall, Defense Manpower Data Center, DoD Center Monterey Bay, 400 Gigling Road, Seaside CA 93955-6771.

0033-3123/96/0600-94210$00.75/0

© 1996 The Psychometric Society

cedures were evaluated according to the Euclidean distance between the true and estimated ability parameters. A method which selects an item that contributes to maximizing the determinant of the Fisher information matrix was judged most precise according to the Euclidean distance criterion.

Missing from previous work (Bloxom & Vale, 1987; Miller, et al., in press; Tam, 1992) however is a theory based procedure for multidimensional adaptive item selection which incorporates prior knowledge of the joint distribution of ability. It will be demonstrated that when the dimensions are correlated, consideration of the prior distribution of ability in item selection can lead to increased measurement efficiency.

Perhaps more importantly, previous work has failed to demonstrate the utility of MAT. This lack of evidence supports the belief by some that MAT offers few benefits, if any, over unidimensional adaptive testing. Wainer et al. (1990) indicate that "although it is certainly theoretically possible to construct a CAT to measure two or more proficiencies simultaneously, it is not yet clear exactly why anyone would want to do that" (p. 242). There are however at least two unique and compelling advantages of MAT.

First, MAT can provide an efficient approach for ensuring adequate coverage of content in adaptive testing. In a test of general science, for example, there may be concern about providing an adequate number of *life*, *physical*, and *chemistry* science items to each examinee, since these tend to be separate (but highly intercorrelated) dimensions of science proficiency. A common approach in unidimensional adaptive testing forces the item selection algorithm to administer a fixed number of items from each content area. However, this approach can be problematic if content is confounded with item difficulty (e.g., if chemistry items are more difficult than life or physical science items). Forcing the administration of chemistry items to an examinee of low ability would provide little information about his level of science proficiency, and would lead to reduced measurement efficiency.

Rather than forcing the item-selection algorithm to administer a fixed number of items from each content area, an alternative approach using MAT would treat these three content areas as separate, but highly intercorrelated dimensions. By incorporating information from several sources on all dimensions simultaneously (including examinee proficiency, item information, and the prior joint-distribution of ability), MAT can in principle provide an efficient choice of items which helps ensure an appropriate coverage of content for each examinee. Although content balancing constraints in unidimensional adaptive testing can result in the administration of items of inappropriate difficulty, MAT item selection procedures can for the most part avoid this problem—providing items of appropriate content coverage and difficulty level.

Increased measurement efficiency provides a second and perhaps more compelling justification for MAT. Increased efficiency can occur when the dimensions measured by the battery have non-zero correlations. In a Bayesian estimation framework, the responses to items of one test can provide information about the level of proficiency measured by the other tests in the battery, provided the p dimensions are correlated. Rather than obtaining p separate ability estimates as in unidimensional adaptive testing, MAT provides a single p-dimensional vector of estimated abilities for each examinee. After each administered item, the p-dimensional vector is updated. For example, in a battery that includes "reading-comprehension" and "vocabulary" subtests, a correct response to a vocabulary item would result in increased provisional estimates for both the vocabulary and reading-comprehension dimensions, since the two are positively correlated. The response to the vocabulary item provides information not only about the vocabulary dimension, but also provides information about the level of reading-comprehension, as well as the level of proficiency on other dimensions that have non-zero correlations with vocabulary. One primary benefit of MAT is that this added information provided by items of correlated dimensions can lead to greater measurement efficiency—manifested by either greater precision or reduced test-lengths. However, it remains to be seen how large this efficiency gain would be under realistic conditions, and whether this gain is large enough to justify the added computational complexities of MAT.

This paper extends previous work (Bloxom & Vale, 1987; Miller, et al., in press; Tam, 1992) in several respects. First a theory-based procedure for item selection (which incorporates prior knowledge of the joint distribution of ability) is presented and evaluated within a multidimensional adaptive framework. In addition, maximum likelihood (ML) and Bayesian-modal ability estimators are presented for the general p-dimensional problem. This paper also presents a comparison of measurement efficiency between MAT and one-dimensional adaptive testing (OAT) using simulated data based on the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). In addition, a discussion of issues relating to the operational use of MAT is provided.

The Item Response Model

We begin by denoting a set of p traits by the vector $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}$. We assume that each of these p traits affects performance on one or more test items. Next we define the item response function (Hattie, 1981) for item i by

$$P_i(\boldsymbol{\theta}) \equiv P(U_i = 1|\boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\mathbf{a}_i'(\boldsymbol{\theta} - b_i\mathbf{1})\right]}, \tag{1}$$

where

- U_i is the binary random variable, containing the response to item i ($U_i = 1$, if item i is answered correctly; and $U_i = 0$, otherwise),
- c_i is the probability that a person with infinitely low ability will answer item i correctly,
- b_i is the difficulty parameter of item i,
- $\mathbf{1}$ is a $p \times 1$ vector of 1's,
- D is the constant 1.7, and
- $\mathbf{a}_i'$ is a $1 \times p$ vector of discrimination parameters for item i.

Note that the exponent in the denominator of (1) can be expressed in scalar notation by

$$-D\mathbf{a}_i'(\boldsymbol{\theta} - b_i\mathbf{1}) = -D\sum_{k=1}^{p} a_{ki}(\theta_k - b_i). \tag{2}$$

We observe that a convenient property of this model occurs when $p = 1$. From the right hand side of (2), we see that (1) reduces to

$$P(U_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i(\theta - b_i)\right]},$$

which is the three parameter logistic (3PL) test model given by Birnbaum (1968).

Another desirable property of the model given by (1) is that the item parameters are unaffected by the distribution of ability. Thus the probability of a correct response depends only on $\mathbf{a}_i$, b_i, c_i and $\boldsymbol{\theta}_0$, and not on the joint distribution of $\boldsymbol{\theta}$.

Note that the item response function given by (1) possesses a single difficulty parameter, rather than separate difficulty parameters for each dimension. Although separate difficulty parameters are conceptually plausible, they are indeterminate and thus cannot be estimated from observed response data.

Throughout this development, we make the assumption of local independence. This assumption states that the probability of a set of observed responses $u_1, u_2, \ldots, u_n$ for an examinee of ability $\boldsymbol{\theta}$ is equal to the product of the probabilities associated with the response to each item

$$P(U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} P_i(\boldsymbol{\theta})^{u_i}(1 - P_i(\boldsymbol{\theta}))^{1-u_i}.$$

This assumption implies that the probability of a correct response to a specific item is a function of only the vector of abilities $\boldsymbol{\theta}$ and the item parameters. Additional information about the probability of a correct response will not be provided from knowledge of performance on any other test item.

Multidimensional adaptive testing, like its unidimensional counterpart requires two complementary procedures. One is an estimation procedure used to obtain a provisional ability estimate after each administered item. The objective of ability estimation is to specify the ability parameters $\boldsymbol{\theta}$ from a set of observed binary responses $\mathbf{u} = \{u_1, u_2, \ldots, u_n\}$ from a single examinee. A second required procedure in MAT is an item selection algorithm which provides an efficient choice of items based on the examinees provisional ability estimate.

Two commonly used methods of ability estimation in unidimensional IRT, maximum likelihood and Bayesian estimation, have direct multidimensional counterparts. Associated with each of the two estimation procedures are methods for quantifying the level of uncertainty in the ability estimates, and adaptive item selection methods.

Maximum Likelihood Estimation and Item Selection

Maximum likelihood estimation begins with a specification of the likelihood function. The likelihood of a vector of observed responses $\mathbf{u}$ given ability $\boldsymbol{\theta}$ is expressed by

$$L(\mathbf{u}|\boldsymbol{\theta}) \equiv L(u_{v_1}, u_{v_2}, \ldots|\boldsymbol{\theta}) = \prod_{i \in \mathbf{v}} P_i(\boldsymbol{\theta})^{u_i} Q_i(\boldsymbol{\theta})^{1-u_i}, \tag{3}$$

where $P_i(\boldsymbol{\theta})$ is defined by (1), $Q_i(\boldsymbol{\theta}) = 1 - P_i(\boldsymbol{\theta})$, and $\mathbf{v}$ is a vector containing the identifiers of the adaptively administered items. The vector of values $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\}$ that maximize the likelihood function given by (3) is taken as the estimator of $\boldsymbol{\theta}$. Tam (1992) has provided expressions based on the normal ogive model for obtaining ML estimators of $\boldsymbol{\theta}$ in the two-dimensional latent space. Here we extend this work to the p-dimensional logistic item response model given by (1).

The ML estimates are the solution to the set of p simultaneous equations given by

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\mathbf{u}|\boldsymbol{\theta}) = \mathbf{0}, \tag{4}$$

where

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\mathbf{u}|\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{\partial}{\partial \theta_1} \ln L(\mathbf{u}|\boldsymbol{\theta}) \\ \dfrac{\partial}{\partial \theta_2} \ln L(\mathbf{u}|\boldsymbol{\theta}) \\ \vdots \\ \dfrac{\partial}{\partial \theta_p} \ln L(\mathbf{u}|\boldsymbol{\theta}) \end{bmatrix}. \tag{5}$$

Explicit expressions for these partial derivatives can be obtained by first noting that the natural logarithm of the likelihood function (3) is

$$\ln L(\mathbf{u}|\boldsymbol{\theta}) = \sum_{i \in \mathbf{v}} [u_i \ln P_i(\boldsymbol{\theta}) + (1 - u_i) \ln Q_i(\boldsymbol{\theta})].$$

The derivative of the log likelihood with respect to θ_k (for $k = 1, 2, \ldots, p$) takes on a form similar to the univariate 3PL model (Lord, 1980, p. 58):

$$\begin{aligned} \frac{\partial}{\partial \theta_k} \ln L(\mathbf{u}|\boldsymbol{\theta}) &= \textstyle\sum_{i \in \mathbf{v}} \left[u_i \dfrac{P_i'(\boldsymbol{\theta})}{P_i(\boldsymbol{\theta})} - (1 - u_i) \dfrac{P_i'(\boldsymbol{\theta})}{Q_i(\boldsymbol{\theta})} \right] \\ &= \textstyle\sum_{i \in \mathbf{v}} \left[[u_i - P_i(\boldsymbol{\theta})] \dfrac{P_i'(\boldsymbol{\theta})}{P_i(\boldsymbol{\theta})Q_i(\boldsymbol{\theta})} \right], \end{aligned} \tag{6}$$

where $P_i'(\boldsymbol{\theta}) = \partial P_i(\boldsymbol{\theta})/\partial \theta_k$. The explicit form for $P_i'(\boldsymbol{\theta})$ is given by

$$\frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_k} = \frac{D a_{ki} Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i]}{1 - c_i}. \tag{7}$$

Substituting (7) into (6) and simplifying, we have

$$\frac{\partial}{\partial \theta_k} \ln L(\mathbf{u}|\boldsymbol{\theta}) = D \sum_{i \in \mathbf{v}} \frac{a_{ki}[P_i(\boldsymbol{\theta}) - c_i][u_i - P_i(\boldsymbol{\theta})]}{(1 - c_i)P_i(\boldsymbol{\theta})}, \tag{8}$$

for $k = 1, 2, \ldots, p$.

Since the likelihood equations (4) have no closed form solutions, an iterative numerical procedure must be used. One popular method is the Newton-Raphson procedure. Let $\boldsymbol{\theta}^{(j)}$ denote the j-th approximation to the value of $\boldsymbol{\theta}$ that maximizes $L(\mathbf{u}|\boldsymbol{\theta})$. Then provided $\boldsymbol{\theta}^{(j)}$ is in the neighborhood of the maximum, an approximation with an even higher likelihood is given by

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \boldsymbol{\delta}^{(j)}, \tag{9}$$

where $\boldsymbol{\delta}^{(j)}$ is the $p \times 1$ vector

$$\boldsymbol{\delta}^{(j)} = [\mathbf{H}(\boldsymbol{\theta}^{(j)})]^{-1} \times \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\mathbf{u}|\boldsymbol{\theta}^{(j)}). \tag{10}$$

The matrix $\mathbf{H}(\boldsymbol{\theta}^{(j)})$ is the $p \times p$ matrix of second derivatives evaluated at $\boldsymbol{\theta}^{(j)}$. The elements of $\mathbf{H}(\boldsymbol{\theta})$ can be expressed by the $p \times p$ symmetric matrix

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{bmatrix} \partial^2 \ln L/\partial\theta_1^2 & \partial^2 \ln L/\partial\theta_1\partial\theta_2 & \cdots & \partial^2 \ln L/\partial\theta_1\partial\theta_p \\ & \partial^2 \ln L/\partial\theta_2^2 & \cdots & \partial^2 \ln L/\partial\theta_2\partial\theta_p \\ & & \ddots & \vdots \\ & & & \partial^2 \ln L/\partial\theta_p^2 \end{bmatrix}.$$

The diagonal elements of $\mathbf{H}(\boldsymbol{\theta})$ take the form

$$\frac{\partial^2}{\partial\theta_k^2} \ln L = D^2 \sum_{i\in\mathbf{v}} \frac{a_{ki}^2 Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1-c_i)^2}, \tag{11}$$

and the off-diagonal elements are of the form

$$\frac{\partial^2}{\partial\theta_k\partial\theta_l} \ln L = D^2 \sum_{i\in\mathbf{v}} \frac{a_{ki}a_{li} Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1-c_i)^2}. \tag{12}$$

In (10), $\partial/\partial\boldsymbol{\theta} \ln L(\mathbf{u}|\boldsymbol{\theta}^{(j)})$ is the $p \times 1$ vector partial derivatives (evaluated at $\boldsymbol{\theta}^{(j)}$) defined by (5). Successive approximations are repeatedly obtained using (9) and (10) until the elements of $\boldsymbol{\delta}^{(j)}$ become sufficiently small.

Note that if the initial values of $\boldsymbol{\theta}^{(j)}$ are not close to the true maximum, then the algorithm given by (9) and (10) may not converge. Convergence can be ensured (at the possible expense of increasing the number of iterations) by using Fisher's method of scoring. This method replaces $\mathbf{H}(\boldsymbol{\theta}^{(j)})$ in (10) with $\mathrm{E}[\mathbf{H}(\boldsymbol{\theta}^{(j)})] = -\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ where $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is the Fisher information matrix, defined by (13). As with the Newton-Raphson method, successive approximations are repeatedly obtained until the elements of $\boldsymbol{\delta}^{(j)}$ become sufficiently small.

Dispersion Matrix

A useful property of ML estimates, denoted by $\hat{\boldsymbol{\theta}}$, is that they tend under regularity conditions to a multivariate normal distribution, with dispersion matrix whose inverse is given by the $p \times p$ information matrix, denoted by $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$. The {$r$-th, s-th} element of this matrix is given by

$$I_{rs}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = -\mathrm{E}\left[\frac{\partial^2 \ln L}{\partial\theta_r\partial\theta_s}\right]. \tag{13}$$

Taking the expectation of (11), we see that the diagonal elements of $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ take the form

$$I_{rr}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = -D^2 \sum\nolimits_{i\in\mathbf{v}} \frac{a_{ri}^2 Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i P_i(\boldsymbol{\theta}) - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1-c_i)^2}$$

$$= \sum\nolimits_{i\in\mathbf{v}} \frac{\left[\dfrac{\partial P_i(\boldsymbol{\theta})}{\partial\theta_r}\right]^2}{P_i(\boldsymbol{\theta})Q_i(\boldsymbol{\theta})}.$$

Similarly, from the expectation of (12) we see that the off-diagonal elements are

$$I_{rs}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = -D^2 \sum\nolimits_{i\in\mathbf{v}} \frac{a_{ri}a_{si} Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i P_i(\boldsymbol{\theta}) - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1-c_i)^2}$$

$$= \sum_{i \in \mathrm{v}} \frac{\partial P_i(\boldsymbol{\theta})/\partial \theta_r \times \partial P_i(\boldsymbol{\theta})/\partial \theta_s}{P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta})}.$$

Note that since each element of the matrix $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is formed from item level summands, we can define an item information matrix, denoted by $\mathbf{I}(\boldsymbol{\theta}, u_i)$, with diagonal elements

$$I_{rr}(\boldsymbol{\theta}, u_i) = \frac{\left[\dfrac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_r}\right]^2}{P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta})}, \tag{14}$$

and off-diagonal elements

$$I_{rs}(\boldsymbol{\theta}, u_i) = \frac{\dfrac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_r} \times \dfrac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_s}}{P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta})}. \tag{15}$$

Adaptive Item Selection

In unidimensional adaptive testing, items can be selected on the basis of item information. The provisional ability estimate $\hat{\theta}_j$ obtained after answering the j-th item is used to evaluate the item information function (Lord, 1980, p. 72):

$$I(\theta, u_{k'}) = \frac{\left[\dfrac{\partial P_{k'}(\hat{\theta}_j)}{\partial \theta}\right]^2}{P_{k'}(\hat{\theta}_j) Q_{k'}(\hat{\theta}_j)},$$

(for $k' \notin \mathrm{v}$). In general, the greatest reduction in the sampling variance of $\hat{\theta}$ is achieved by administering the item with the largest information value.

A multivariate item-selection analog is motivated by the expression for the volume of the multivariate normal ellipsoid (Anderson, 1984, p. 263). Since the provisional ability estimate $\hat{\boldsymbol{\theta}}_j$ (obtained from the first j responses) is distributed asymptotically according to a multivariate normal distribution $N(\hat{\boldsymbol{\theta}}_0, \boldsymbol{\Sigma}_j)$, then

$$\Pr[\hat{\boldsymbol{\theta}}_j' \, \boldsymbol{\Sigma}_j^{-1} \hat{\boldsymbol{\theta}}_j \le \chi_p^2(\alpha)] = 1 - \alpha.$$

That is, the probability is $1 - \alpha$ that $\hat{\boldsymbol{\theta}}_j$ will fall inside the ellipsoid

$$\mathbf{x}' \boldsymbol{\Sigma}_j^{-1} \mathbf{x} = \chi_p^2(\alpha).$$

The volume of this ellipsoid is

$$\varsigma \times |\boldsymbol{\Sigma}_j|^{1/2},$$

where

$$\varsigma = \frac{2\pi^{p/2}[\chi_p^2(\alpha)]^{p/2}}{p\Gamma(\frac{1}{2}p)}, \tag{16}$$

and $\Gamma(\cdot)$ denotes the gamma function.

When considering items for administration, the multivariate analog to the univariate procedure selects the item that achieves the largest decrement in the volume of the confidence ellipsoid. We denote the volume decrement achieved by the administration of item k' by

$$V_{k'} = \varsigma|\boldsymbol{\Sigma}_j|^{1/2} - \varsigma|\boldsymbol{\Sigma}_{j+k'}|^{1/2}, \tag{17}$$

where $\boldsymbol{\Sigma}_j$ is the dispersion matrix of the $p \times 1$ vector of provisional estimates $\hat{\boldsymbol{\theta}}_j$ obtained after the j-th response, and $\boldsymbol{\Sigma}_{j+k'}$ is the dispersion matrix of provisional estimates obtained after administration of the first j items and the administration of item k'. Since $\hat{\boldsymbol{\theta}}_j$ are ML estimates, $\boldsymbol{\Sigma}_j$ can be approximated by the inverse of the information matrix, given by

$$\boldsymbol{\Sigma}_j = \{\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)\}^{-1} = \left[\sum_{i \in \mathrm{v}} \mathbf{I}(\boldsymbol{\theta}, u_i)\right]^{-1}. \tag{18}$$

The covariance matrix of provisional estimates which includes the administration of item k' is given by

$$\boldsymbol{\Sigma}_{j+k'} = [\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) + \mathbf{I}(\boldsymbol{\theta}, u_{k'})]^{-1}. \tag{19}$$

Substituting (18) and (19) into (17), we obtain

$$V_{k'} = \varsigma|[\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)]^{-1}|^{1/2} - \varsigma|[\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) + \mathbf{I}(\boldsymbol{\theta}, u_{k'})]^{-1}|^{1/2}.$$

The expression for the volume decrement can be further simplified by noting that the determinant of the inverse of $\mathbf{I}$ is the reciprocal of the determinant (Searle, 1982, p. 130):

$$V_{k'} = \varsigma|\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)|^{-1/2} - \varsigma|\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) + \mathbf{I}(\boldsymbol{\theta}, u_{k'})|^{-1/2}.$$

Note that the first term is a constant across items, since ς depends only on p and α, and $|\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)|$ is based on previously administered items. Since in the second term ς remains constant over candidate items, $V_{k'}$ can be maximized by selecting the item that maximizes the quantity

$$|\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) + \mathbf{I}(\boldsymbol{\theta}, u_{k'})|. \tag{20}$$

The ML approach to estimation and item selection has two undesirable qualities. First, towards the beginning of the adaptive test, the ability estimates contained in the p-element vector $\hat{\boldsymbol{\theta}}_j$ will be either undefined or poorly defined. Consequently, some heuristic procedure is needed to define particular elements in $\hat{\boldsymbol{\theta}}_j$ in the absence of sufficient data. Second, the ML item selection based on the volume of the confidence ellipsoid does not consider prior knowledge about the joint distribution of $\boldsymbol{\theta}$. However, these shortcommings can be remedied by applying Bayesian methodology to the problems of item selection and ability estimation.

Bayesian Estimation and Item Selection

According to Bayes theorem, the posterior density function of $\boldsymbol{\theta}$ is expressed by

$$f(\boldsymbol{\theta}|\mathbf{u}) = L(\mathbf{u}|\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{f(\mathbf{u})}, \tag{21}$$

where $L(\mathbf{u}|\boldsymbol{\theta})$ is the likelihood function given by (3), $f(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, and $f(\mathbf{u})$ is the marginal probability of $\mathbf{u}$. Here, we shall consider the case in which the prior distribution of $\boldsymbol{\theta}$ is multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Phi}$:

$$f(\boldsymbol{\theta}) = (2\pi)^{-p/2}|\boldsymbol{\Phi}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right].$$

Once the prior density has been specified, then the posterior density contains all necessary information about $\boldsymbol{\theta}$. However, in the form given by (21), this information is not readily usable. Point estimates of ability are usually defined as the mean or the mode of the posterior distribution. Since the mode of the posterior distribution (modal estimate) requires far fewer computations than the posterior mean for problems of higher dimensionality, we focus on its application to the problem of multidimensional adaptive estimation.

The modal estimates are those values of the parameters that correspond to the maximum of the posterior density function. These can be obtained by maximizing the natural logarithm of the posterior distribution. The modal estimates, denoted by $\boldsymbol{\theta}^*$, are the values of $\boldsymbol{\theta}$ that satisfy the set of p simultaneous equations given by

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}|\mathbf{u}) = \mathbf{0},$$

where

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}|\mathbf{u}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln f(\boldsymbol{\theta}|\mathbf{u}) \\ \frac{\partial}{\partial \theta_2} \ln f(\boldsymbol{\theta}|\mathbf{u}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln f(\boldsymbol{\theta}|\mathbf{u}) \end{bmatrix}. \tag{22}$$

Explicit expressions for these partial derivatives can be obtained by noting that the natural logarithm of the posterior density function (21) is

$$\begin{aligned} \ln f(\boldsymbol{\theta}|\mathbf{u}) &= \ln L(\mathbf{u}|\boldsymbol{\theta}) + \ln f(\boldsymbol{\theta}) + \text{constant} \\ &= \ln L(\mathbf{u}|\boldsymbol{\theta}) - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) + \text{constant}. \end{aligned}$$

Then we have

$$\frac{\partial \ln f(\boldsymbol{\theta}|\mathbf{u})}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \ln L(\mathbf{u}|\boldsymbol{\theta}) - \frac{1}{2}\frac{\partial}{\partial \theta_k}[(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})]. \tag{23}$$

Expressions for the first term, $\partial \ln L(\mathbf{u}|\boldsymbol{\theta})/\partial \theta_k$, are provided by (8). The explicit expression for the second term takes the form

$$\frac{\partial}{\partial \theta_k}[(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})] = 2 \times \left[\frac{\partial}{\partial \theta_k}(\boldsymbol{\theta} - \boldsymbol{\mu})'\right]\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}), \tag{24}$$

where

$$\frac{\partial}{\partial \theta_k}(\boldsymbol{\theta} - \boldsymbol{\mu})'$$

denotes a $1 \times p$ vector with the k-th element set equal to 1 and all other elements equal to zero. Substituting (8) and (24) into (23), we have

$$\frac{\partial \ln f(\boldsymbol{\theta}|\mathbf{u})}{\partial \theta_k} = D \sum_{i \in \mathrm{v}} \frac{a_{ki}[P_i(\boldsymbol{\theta}) - c_i][u_i - P_i(\boldsymbol{\theta})]}{(1 - c_i)P_i(\boldsymbol{\theta})} - \left[\frac{\partial}{\partial \theta_k}(\boldsymbol{\theta} - \boldsymbol{\mu})'\right]\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) \tag{25}$$

(for $k = 1, 2, \ldots, p$).

As with the likelihood equations, the equations given by (25) have no explicit solutions, so an iterative numerical procedure such as the Newton-Raphson procedure must be used. Accordingly, if we let $\boldsymbol{\theta}^{(j)}$ denote the j-th approximation to the value of $\boldsymbol{\theta}$ that maximizes $\ln f(\boldsymbol{\theta}|\mathbf{u})$, then a better approximation is generally given by

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \boldsymbol{\delta}^{(j)}, \tag{26}$$

where $\boldsymbol{\delta}^{(j)}$ is the $p \times 1$ vector

$$\boldsymbol{\delta}^{(j)} = [\mathbf{J}(\boldsymbol{\theta}^{(j)})]^{-1} \times \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}^{(j)}|\mathbf{u}). \tag{27}$$

The matrix $\mathbf{J}(\boldsymbol{\theta}^{(j)})$ is the $p \times p$ matrix of second derivatives evaluated at $\boldsymbol{\theta}^{(j)}$. The elements of $\mathbf{J}(\boldsymbol{\theta})$ are expressed by the $p \times p$ symmetric matrix

$$\mathbf{J}(\boldsymbol{\theta}) = \begin{bmatrix} \partial^2 \ln f(\boldsymbol{\theta}|\mathbf{u})/\partial \theta_1^2 & \partial^2 \ln f(\boldsymbol{\theta}|\mathbf{u})/\partial \theta_1 \partial \theta_2 & \cdots & \partial^2 \ln f(\boldsymbol{\theta}|\mathbf{u})/\partial \theta_1 \partial \theta_p \\ & \partial^2 \ln f(\boldsymbol{\theta}|\mathbf{u})/\partial \theta_2^2 & \cdots & \partial^2 \ln f(\boldsymbol{\theta}|\mathbf{u})/\partial \theta_2 \partial \theta_p \\ & & \ddots & \vdots \\ & & & \partial^2 \ln f(\boldsymbol{\theta}|\mathbf{u})/\partial \theta_p^2 \end{bmatrix}.$$

Taking the derivative of (23), we see that the diagonal elements of $\mathbf{J}(\boldsymbol{\theta})$ take the form

$$\frac{\partial^2}{\partial \theta_k^2} \ln f(\boldsymbol{\theta}|\mathbf{u}) = \frac{\partial^2}{\partial \theta_k^2} \ln L(\mathbf{u}|\boldsymbol{\theta}) - \frac{1}{2}\frac{\partial^2}{\partial \theta_k^2}[(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})]. \tag{28}$$

The first term on the right hand side of (28) is given by (11). The explicit expression for the second term is given by

$$\frac{\partial^2}{\partial \theta_k^2}[(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})] = 2 \times \left[\frac{\partial}{\partial \theta_k}(\boldsymbol{\theta} - \boldsymbol{\mu})'\right]\boldsymbol{\Phi}^{-1}\left[\frac{\partial}{\partial \theta_k}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]. \tag{29}$$

Substituting (11) and (29) into (28) we have

$$\frac{\partial^2}{\partial \theta_k^2} \ln f(\boldsymbol{\theta}|\mathbf{u}) = D^2 \sum_{i \in \mathrm{v}} \frac{a_{ki}^2 Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1 - c_i)^2} - \phi^{kk}, \tag{30}$$

where ϕ^{kk} is the k-th diagonal element of $\boldsymbol{\Phi}^{-1}$. The off-diagonal elements of $\mathbf{J}(\boldsymbol{\theta})$ take the form

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} \ln f(\boldsymbol{\theta}|\mathbf{u}) = \frac{\partial^2}{\partial \theta_k \partial \theta_l} \ln L(\mathbf{u}|\boldsymbol{\theta}) - \frac{1}{2}\frac{\partial^2}{\partial \theta_k \partial \theta_l}[(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})].$$

From (12) and by evaluating the second term we have

$$\frac{\partial^2}{\partial\theta_k\partial\theta_l}\ln f(\boldsymbol{\theta}|\mathbf{u}) = D^2\sum_{i\in\mathbf{v}}\frac{a_{ki}a_{li}Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_iu_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1 - c_i)^2} - \phi^{kl}, \tag{31}$$

where ϕ^{kl} is the $\{k\text{-th}, l\text{-th}\}$ element of $\boldsymbol{\Phi}^{-1}$. The vector of elements $\partial\ln f(\boldsymbol{\theta}^{(j)}|\mathbf{u})/\partial\boldsymbol{\theta}$ in (27) is the $p \times 1$ vector of partial derivatives (evaluated at $\boldsymbol{\theta}^{(j)}$) defined by (25). Modal estimates can be obtained through successive approximations using (26) and (27). Additional approximations are obtained until the elements of $\boldsymbol{\theta}^{(j)}$ change very little from one iteration to the next.

If the initial values of $\boldsymbol{\theta}^{(j)}$ are not in the neighborhood of the maximum, then the iterative procedure given by (26) and (27) may not converge. This problem can be avoided by using Fisher's method of scoring, where $\mathrm{E}[\mathbf{J}(\boldsymbol{\theta}^{(j)})] = -\mathbf{W}$ is substituted for $\mathbf{J}(\boldsymbol{\theta}^{(j)})$ in (27). The matrix $\mathbf{W}$ has special application to item selection and its derivation is presented below.

Adaptive Item Selection

Here we consider a Bayesian approach to multidimensional item selection, where the next item is chosen to provide the largest decrement in the volume of the credibility ellipsoid. For a normal posterior density function, the volume decrement achieved by the administration of item k' is given by

$$C_{k'} = \zeta|\mathbf{W}_j^{-1}|^{1/2} - \zeta|\mathbf{W}_{j+k'}^{-1}|^{1/2},$$

where $\mathbf{W}_j^{-1}$ is the covariance matrix of the posterior distribution computed from the first j items, $\mathbf{W}_{j+k'}^{-1}$ is the covariance matrix incorporating $j + 1$ items (the first j items plus item k'), and ζ is defined by (16). For the purpose of item selection, we approximate the posterior density function $f(\boldsymbol{\theta}|\mathbf{u})$ by a multivariate normal density with covariance matrix $\mathbf{W}^{-1}$, where

$$\mathbf{W} = -\mathrm{E}[\mathbf{J}(\boldsymbol{\theta})],$$

and where $-\mathrm{E}[\mathbf{J}(\boldsymbol{\theta})]$ is evaluated at the mode of the posterior distribution $\boldsymbol{\theta}^*$. Taking the expectation of (30), we see that the diagonal elements of $\mathbf{W}$ take the form

$$w_{rr} = -D^2\sum_{i\in\mathbf{v}}\frac{a_{ri}^2Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_iP_i(\boldsymbol{\theta}) - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1 - c_i)^2} + \phi^{rr}, \tag{32}$$

while the off-diagonal elements from (31) are

$$w_{rs} = -D^2\sum_{i\in\mathbf{v}}\frac{a_{ri}a_{si}Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_iP_i(\boldsymbol{\theta}) - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1 - c_i)^2} + \phi^{rs}. \tag{33}$$

The matrix $\mathbf{W}_j$ is computed from (32) and (33) where the summands are taken over the j adaptively administered items $\mathbf{v} = \{v_1, v_2, \ldots, v_j\}$, whereas the matrix $\mathbf{W}_{j+k'}$ is computed from the summands $\mathbf{v} = \{v_1, v_2, \ldots, v_j, v_{k'}\}$.

The expression for the volume decrement can be simplified by noting that the determinant of the inverse of $\mathbf{W}$ is the reciprocal of the determinant (Searle, 1982, p. 130):

$$C_{k'} = \varsigma|\mathbf{W}_j|^{-1/2} - \varsigma|\mathbf{W}_{j+k'}|^{-1/2}. \tag{34}$$

Note that the first term is a constant across candidate items, since ς depends only on p and α, and $|\mathbf{W}_j|$ is based on previously administered items. The second term is a

TABLE 1

CAT-ASVAB Characteristics

Subtest	Content Area	Test-Length	Pool-Size
1	General Science (GS)	15	110
2	Arithmetic Reasoning (AR)	15	209
3	Word Knowledge (WK)	15	228
4	Paragraph Comprehension (PC)	10	88
5	Auto Information (AI)	10	104
6	Shop Information (SI)	10	103
7	Math Knowledge (MK)	15	103
8	Mechanical Comprehension (MC)	15	103
9	Electronics Information (EI)	15	97

function of both ς and the determinate of the matrix $\mathbf{W}_{j+k'}$. Since ς remains constant over candidate items, $C_{k'}$ can be maximized by selecting the item k' which maximizes $|\mathbf{W}_{j+k'}|$. We can note the relation between $|\mathbf{W}_{j+k'}|$ and the criterion used in the ML procedure (20) from the equation

$$|\mathbf{W}_{j+k'}| = |\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) + \mathbf{I}(\boldsymbol{\theta}, u_{k'}) + \boldsymbol{\Phi}^{-1}|. \tag{35}$$

The equivalence of the right and left hand sides of (35) can be easily verified from the definitions of $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ and $\mathbf{I}(\boldsymbol{\theta}, u_{k'})$ given by (13), (14) and (15), and from the definition of **W** given by (32) and (33). Note that the criterion for the ML item selection (20) and the criterion for the Bayesian item selection based on $|\mathbf{W}_{j+k'}|$ differ only by the term which consists of the inverse of the covariance matrix of the prior distribution of abilities $\boldsymbol{\Phi}^{-1}$.

Simulation Study

Reliability values for the multidimensional Bayesian ability estimates were compared to their unidimensional counterparts using simulated test sessions. In addition, the relative contribution of the multivariate item selection and scoring algorithms to the increased efficiency of MAT was also examined. The simulated tests were based on the nine adaptive power tests of the CAT-ASVAB (Segall, Moreno, & Hetter, 1987). For these simulations, item parameters of Form 1 of the CAT-ASVAB were used. The nine tests, along with test-lengths and pool sizes are listed in Table 1. Form 1 consists of 1145 items which span nine content areas.

Item parameter estimates used in the operational CAT-ASVAB were treated as population values in the simulation study. Although the parameter estimates were obtained using the unidimensional 3PL model, they were adapted for use in the multidimensional model given by (1). A nine dimensional model was used, one dimension for each of the nine content areas contained in the CAT-ASVAB. Each of the nine elements contained in the vector of discrimination parameters $\mathbf{a}_i'$ corresponded to a different content area. Each item was allowed to possess one nonzero discrimination parameter. Using this convention, the position of the nonzero element corresponds to the content area of the item. For example, the vectors of discrimination parameters for GS items were of the form $\mathbf{a}_i' = \{a_{1i}, 0, 0, 0, 0, 0, 0, 0, 0\}$, where $a_{1i} > 0$. Similarly, the vectors of discrimination parameters for AR items (the second test in the battery) took the form $\mathbf{a}_i' = \{0, a_{2i}, 0, 0, 0, 0, 0, 0, 0\}$, and so forth for the remaining seven

TABLE 2

Covariance Matrix of Latent Abilities Φ

	Dimension								
Dimension	GS	AR	WK	PC	AI	SI	MK	MC	EI
GS	1.000								
AR	.645	1.000							
WK	.908	.611	1.000						
PC	.808	.847	.880	1.000					
AI	.486	.332	.326	.349	1.000				
SI	.676	.424	.566	.514	.824	1.000			
MK	.564	.846	.516	.711	.150	.218	1.000		
MC	.739	.758	.644	.800	.623	.725	.625	1.000	
EI	.808	.639	.724	.743	.642	.724	.536	.822	1.000

content areas. From a factor analytic viewpoint, this formulation provides a simple structure, where each item loads on a single dimension. The values of the difficulty, guessing, and non-zero discrimination parameters for each item corresponded to their estimated values obtained in the unidimensional 3PL calibration (Prestwood, Vale, Massey, & Welsh, 1985).

The population covariance matrix Φ was specified from the disattenuated correlation matrix of the nine adaptive power tests contained in the CAT-ASVAB. These disattenuated correlations were estimated from data obtained in an alternate forms reliability study of the CAT-ASVAB (Moreno & Segall, 1992). The covariance matrix of latent abilities Φ is given in Table 2.

Multidimensional Simulations

A total of 15 conditions were simulated using the MAT Bayesian item selection and scoring algorithm. These conditions varied in the total number of items administered to each simulated examinee. These ranged from a low of 9 items (for Condition MAT-9) to a high of 120 items (for Condition MAT-120). Test-lengths for each condition are provided in Table 3. Each of the 15 conditions provided restrictions on the total number of items administered from each of the nine content areas. For example, as indicated in Table 3, only one item was administered from each of the nine content areas in the first condition (MAT-9). In the second condition (MAT-17), one item was administered from PC and SI, three items from MK, and two items were administered from each of the remaining six content areas. Note that the proportion of items administered from each content area in the MAT simulation differs from the corresponding proportions administered in the simulation of the unidimensional battery (as indicated by a comparison of Conditions MAT-120 and OAT in Table 3). In the multidimensional simulations, the proportion of items for AI and MK are larger, while the proportion of items from GS, PC, MC, and EI are slightly smaller. These proportions of items provide a pattern of reliabilities that is similar across the MAT and OAT approaches, which facilitates comparisons at different battery lengths. The dependent measures in each condition were reliability values for the nine content areas, estimated from the squared correlations between the true and estimated ability parameters.

Several steps were involved in the simulation process. A sample of 1000 vectors of true abilities were generated from a multivariate normal distribution with mean vector

TABLE 3

Simulation Study Test-Lengths

	Number of Administered Items									
Condition	GS	AR	WK	PC	AI	SI	MK	MC	EI	Total
MAT-9	1	1	1	1	1	1	1	1	1	9
MAT-17	2	2	2	1	2	1	3	2	2	17
MAT-24	2	3	3	2	2	2	4	3	3	24
MAT-32	3	4	4	2	3	3	5	4	4	32
MAT-40	4	5	5	3	4	3	6	5	5	40
MAT-50	5	6	6	4	5	4	8	6	6	50
MAT-58	6	7	7	4	6	5	9	7	7	58
MAT-62	6	8	8	5	6	5	10	7	7	62
MAT-70	7	9	9	5	7	6	11	8	8	70
MAT-80	8	10	10	6	8	7	13	9	9	80
MAT-88	9	11	11	7	9	7	14	10	10	88
MAT-96	10	12	12	7	10	8	15	11	11	96
MAT-103	10	13	13	8	10	9	16	12	12	103
MAT-111	11	14	14	8	11	9	18	13	13	111
MAT-120	12	15	15	9	12	10	19	14	14	120
OAT	15	15	15	10	10	10	15	15	15	120

$\mathbf{0}$ and covariance matrix $\mathbf{\Phi}$. These vectors were generated using the IMSL subroutines CHFAC and RNMVN. The provisional ability estimates for each simulated examinee were initialized to zero. Using the criterion given by (35), the item providing the largest decrement in the volume of the credibility ellipsoid was selected for administration. Simulated responses to selected items were obtained by evaluating the item response function (1) at the true ability level and comparing the probability value to a pseudo random uniform number. After each response, a vector of provisional Bayesian modal ability estimates were obtained using the Newton-Raphson procedure given by (26). Only items that had not been previously administered to the simulated examinee were considered for administration. An additional requirement was that the item must belong to a content area that had not reached its maximum number of administered items. The process of item-selection and provisional ability estimation was repeated for each simulated examinee until the target test-length had been reached for the condition. For each condition (MAT-9 through MAT-120) the squared correlation between the true abilities and final modal estimates were calculated (see Table 4). As indicated, these squared correlations were calculated separately for each of the nine content areas.

Unidimensional Simulations

For comparison purposes, an additional simulation was performed for the OAT approach. This simulation used the same item pools (Table 1) and item parameters used in the MAT simulations. A sample of 2000 vectors of true abilities were generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Phi}$. These vectors were generated using the IMSL subroutines CHFAC and RNMVN. Simulations were conducted separately for each of the nine content areas. Each simulated examinee's initial ability estimate was set to zero. Items were selected on the basis of maximum item information (evaluated at the provisional ability estimate). Responses

TABLE 4

Simulated Reliability Estimates

	Content Area								
Condition	GS	AR	WK	PC	AI	SI	MK	MC	EI
MAT-9	.719	.675	.676	.712	.588	.645	.624	.685	.679
MAT-17	.800	.785	.762	.798	.683	.717	.791	.780	.776
MAT-24	.837	.844	.818	.865	.702	.770	.840	.826	.826
MAT-32	.863	.870	.864	.882	.785	.846	.871	.854	.841
MAT-40	.884	.897	.885	.900	.838	.835	.896	.875	.867
MAT-50	.904	.909	.905	.911	.852	.867	.908	.882	.874
MAT-58	.911	.913	.919	.915	.888	.885	.923	.894	.885
MAT-62	.915	.927	.931	.929	.877	.871	.929	.896	.889
MAT-70	.915	.923	.929	.928	.894	.880	.939	.907	.890
MAT-80	.914	.933	.939	.940	.899	.899	.938	.911	.897
MAT-88	.929	.935	.945	.934	.915	.893	.949	.916	.896
MAT-96	.931	.937	.949	.943	.926	.910	.951	.919	.894
MAT-103	.933	.944	.952	.945	.927	.920	.949	.922	.899
MAT-111	.936	.946	.957	.943	.928	.916	.956	.926	.916
MAT-120	.934	.949	.954	.948	.933	.922	.956	.927	.909
OAT-UME	.904	.928	.938	.856	.902	.876	.943	.876	.879
OAT-MME	.938	.938	.953	.939	.914	.916	.944	.919	.914

were generated by evaluating the 3PL item response function at the examinees true ability and comparing the value to a pseudo random uniform number. The unidimensional provisional ability estimate was updated by setting it equal to the mode of the posterior distribution. This was done using a Newton-Raphson procedure which assumed a normal (0, 1) prior. The item-selection and ability-updating process was repeated for each simulated examinee until the target test-length had been reached.

Two methods of computing final scores were applied to the data of each simulated examinee: (a) the unidimensional modal estimator (UME), and (b) the multidimensional modal estimator (MME). (These estimates were obtained at the end of the adaptive test and did not influence item selection.) The unidimensional modal estimator was calculated by setting the ability estimate equal to the mode of the univariate posterior distribution, for each of the nine content areas. The squared correlations between the 2000 true and estimated abilities were calculated for each of the nine content areas (see Condition "OAT-UME" in Table 4). The multidimensional modal estimator (containing a vector of nine estimated abilities) was obtained using the Newton-Raphson procedure given by (26). The squared correlations between the 2000 true and estimated abilities were calculated for each content area (see Condition "OAT-MME" in the last row in Table 4).

Results

Scatterplots between true and estimated ability parameters for conditions MAT-9, MAT-40, MAT-80, MAT-120, and OAT-UME were generated for each of the nine content areas to verify the assumption of linearity. No departures from linearity were evident, indicating that the squared correlations are suitable measures of reliability.

The reliability values are presented in Table 4. As indicated, the reliability values

TABLE 5

Residual Correlation Matrix: $(\hat{\mathbf{R}} - \boldsymbol{\Phi}) \times 1000$

	Dimension								
Dimension	GS	AR	WK	PC	AI	SI	MK	MC	EI
GS	—	030	037	054	049	052	011	050	060
AR	-077	—	031	031	053	032	027	040	052
WK	-077	-040	—	022	055	050	021	055	062
PC	-090	-110	-085	—	063	057	034	044	066
AI	-038	-011	-009	-002	—	056	016	062	066
SI	-057	-047	-040	-049	-083	—	000	053	076
MK	-049	-062	-018	-077	-015	-024	—	015	023
MC	-068	-089	-055	-113	-072	-074	-062	—	063
EI	-066	-051	-053	-075	-047	-053	-043	-097	—

Note. Below the diagonal: OAT-UME. Above the diagonal: MAT-120.

for each content-area tend to increase as a function of battery test length. The principal finding is made by a comparison of the reliability values of the OAT-UME approach with those of the MAT approach—specifically with MAT-80. The OAT-UME approach (which administered 120 items) has reliability values less than or about equal to those of MAT-80, in which a total of 80 items were administered. Consequently, MAT achieved greater or comparable precision with one-third fewer items.

The second significant finding of the simulations is observed from a comparison of the reliability values of MAT-120 with those of the OAT-UME simulation. Here we observe that with equivalent numbers items across the MAT and OAT-UME approach (120 items), MAT obtains substantially higher reliability values for most of the nine content areas.

This higher reliability obtained by MAT can be attributed to increased efficiencies in: (a) item selection, and (b) ability estimation. Some insight into the relative contributions of the MAT scoring and item selection algorithms can be obtained by examining the reliabilities of the OAT-MME condition. The reliability of one-dimensional adaptive testing can be increased substantially by applying the multidimensional Bayesian scoring algorithm to the complete response pattern (as seen by a comparison of the OAT-UME and OAT-MME conditions in Table 4). From a comparison of the last three conditions in Table 4, it is evident that a substantial portion of the gain in reliability achieved by MAT (over OAT-UME) can be attributed to the use of the multidimensional Bayesian modal estimator—with the remaining gain in precision attributed to the multidimensional item selection algorithm.

The 9 × 9 matrix of correlations among the observed ability estimates $\hat{\mathbf{R}}$ for both the MAT-120 and OAT-UME conditions was calculated and compared to the population matrix $\boldsymbol{\Phi}$. Table 5 provides the residual correlation matrix $\hat{\mathbf{R}} - \boldsymbol{\Phi}$. As indicated, the two procedures provide estimated correlations with bias in opposite directions. The OAT-UME procedure provides estimated correlations that are consistently lower than the population values. This attenuation is most likely due to measurement error. The MAT procedure provides estimated correlations that are consistently higher than the population values. This inflation is most likely due to the bias in the estimates introduced by the prior covariance matrix. Because of these biases, caution should be exercised when making inferences about population correlations from values calculated from observed ability estimates. If these correlations are of interest, then a preferable approach would be to estimate these parameters directly from examinee item responses using a procedure given by Mislevy (1984).

In interpreting the results of the simulation analyses, it is useful to note that there is one special case of MAT that is functionally equivalent to OAT methodology. This case occurs when the items possess simple structure (load on only one dimension) and the prior distribution of abilities $\mathbf{\Phi}$ is a diagonal matrix. In this case the solution to the p equations given by (22) can be solved for each θ_k (for $k = 1, 2, \ldots, p$) separately. These equations reduce to a form that is identical to those used in unidimensional Bayesian modal estimation. In addition it can be shown that the item selection criterion given by (35) will provide an identical rank ordering of items as that provided by the evaluation of the univariate ML item information function. Thus when each item loads on a single dimension and the dimensions are uncorrelated, the MAT item selection and scoring procedures are equivalent to the univariate Bayesian modal estimator used in conjunction with maximum information item selection. One logical inference derived from this relation is that the gains in efficiency obtained by MAT depend on the correlations among the dimensions contained in $\mathbf{\Phi}$. In general, the larger the magnitude of these correlations, the higher the gains in efficiency over OAT.

Issues in Multidimensional Adaptive Testing

The results of the simulation study provide compelling evidence for one of the principal benefits of MAT—that of greater measurement efficiency. This gain in efficiency can be manifested by either a reduction in test-length or by greater measurement precision. The results indicate that MAT can achieve equivalent or higher levels of precision with about one-third fewer items than OAT. The results also indicate that when the total number of items is held constant across the MAT and OAT approaches, MAT can provide a substantial increase in reliability. There are however several refinements and unique applications of MAT that can provide additional benefits. These include the use of MAT to control the balance of item content among examinees, and further increases in efficiency influenced by the choice of alternative item selection and test-termination strategies. However, before MAT can be used on a routine basis, several issues require further consideration and investigation. These areas include item parameter estimation, exposure control, developing a common metric and orientation, and the affect of item-order on item functioning.

Content Balancing

In addition to providing greater measurement efficiency, MAT can also provide a mechanism for addressing issues of content balancing in computerized adaptive testing. One concern in CAT is that in general examinees do not receive equal numbers of items from important content areas—resulting in ability estimates based on different mixtures of contents. For example, a test of general science may consist of items from three content areas: (a) biology, (b) physical sciences, and (c) chemistry. One approach to content balancing in CAT is to allocate the presented items among the content areas in a way that ensures that each examinee receives a specified number of items from each content area. For example we may specify that each examinee receive 5 biological, 5 physical, and 5 chemistry items. This approach to content balancing has an undesirable feature—the items administered from one or more content areas may be uninformative about the general level of proficiency for the examinee. That is, some items may be of inappropriate difficulty level. For example if the chemistry items were difficult relative to the other content areas, then administering these items to examinees of low or moderate ability may provide little additional information about their level of scientific knowledge.

An alternative approach to content balancing using MAT treats each of the three content areas as separate dimensions. Rather than fixing the number of items to be administered from each content area, items are selected based on the multivariate Bayesian item selection criterion given by (35). Using this criterion, the level of proficiency on one dimension is used to help select informative items from the other two dimensions. At each step in the item selection process, the item which reduces the volume of the credibility ellipsoid by the largest amount is selected. For example, low or moderate ability examinees would likely receive primarily biological and physical science items, since the difficult chemistry items provide only small decrements in the volume of the credibility ellipsoid. On the other hand, high ability examinees would not only receive biological and physical science items, but also receive a number of chemistry items, since these items reduce the level of uncertainty about the chemistry dimension for examinees at high ability levels. Using MAT, the balance of content is based on the estimated level of proficiency—examinees at different levels receive an appropriately tailored mixture of item content.

Note that since MAT provides separate ability estimates for each dimension, these values must be combined to form a single composite measure. In unidimensional CAT involving multiple content areas, there is an implicit weighting of each dimension resulting from content restrictions which are placed on item selection. In MAT these weights must be specified explicitly. Once specified, these values can be used to form a weighted linear composite of the separate ability estimates provided for each dimension.

Item Selection and Test Termination Strategies

Within the general approach to multidimensional adaptive testing presented here, there are several constraints that can be placed on item selection and test termination. One possible stopping rule consists of a variable-length strategy where a desired level of precision is specified. Testing proceeds until the volume of the credibility ellipsoid is reduced to the desired value. Although this approach has the desirable property of achieving similar levels of precision across different levels of ability in the multidimensional space, it is likely to suffer from some of the same drawbacks suffered by its unidimensional counterpart. There may be some points in the multidimensional space in which very few items contain adequate levels of information. Examinees which fall in these regions may receive very long tests, with each new item providing very little additional information. This problem can be avoided by placing a maximum on the total number of administered items. According to this strategy, testing would continue until one of the following occurred: (a) the target level of precision was achieved or (b) the maximum number of items was reached. Placing a maximum limit on the number of administered items can help increase measurement efficiency by restricting the test lengths of examinee's who might otherwise receive a substantial number of uninformative items. For a given item pool, the specification of maximum test-lengths can be investigated through a series of MAT simulations. Using simulated administrations, the distribution of test-lengths can be examined for a target precision-level. Based in this distribution, the maximum limit can be chosen to eliminate the occurrence of extraordinarily long tests, while leaving the test-lengths for the majority of examinees unaffected.

Fixed-length testing provides an alternative stopping-rule, where the total number of administered items is fixed for each examinee. For example, we might fix the overall test-length to some specific value, and allow the numbers of items administered from each content area to vary across examinees. Although this approach selects items in the most efficient manner, it may not emphasize important content areas. For example, we may desire a particular level of precision for math ability. If math items are less informative than items of other content areas, the MAT algorithm is likely to administer primarily items of non-math content. Consequently, to achieve the desired level of precision for particular content areas, restrictions on the choice of item content may be required.

One way to achieve a desired level of precision for particular dimensions is to place constraints on the total number of items administered from each content area. In general, a larger number of items would be administered from dimensions that required higher precision. Higher levels of precision might be required of a dimension used for selection or classification decisions. Conversely, lower precision levels might be satisfactory for dimensions which are always combined with others to form composites, upon which decisions are made. If for example math ability alone formed the basis of an important selection decision, we could fix the number of items administered from the math content area at 15, while limiting the number of administered items to 10 for each of the remaining content areas. Note that we would not constrain the order of item presentation. Using this approach the most informative items are selected from among those not already administered, subject to the requirement that the target number of items from the chosen content area has not been reached. Constraints placed in this manner could be used to adjust the relative level of precision for different dimensions of the battery.

Another approach to increased efficiency incorporates expected response latency into the choice of items. By doing so we can select the item which provides the largest decrement in the volume of the credibility ellipsoid "per unit of time". This can be accomplished, for example, by using the index $\rho_{k'} = C_{k'}/t_{k'}$, where $C_{k'}$ is defined by (34), and $t_{k'}$ is the expected response time for item k'. By averaging across examinees to obtain an "expected response time" for individual items, we treat time as a dimension along which items (rather than examinees) are ordered. Since the expected response time can vary greatly across items of different content areas (and among items within a content area), the rank-ordering of candidate items may vary greatly across the $C_{k'}$ and $\rho_{k'}$ indices. For example, the average response time is generally much shorter for vocabulary items than for paragraph comprehension items, even though the dimensions measured by the two content areas are highly correlated. Using the ratio criterion $\rho_{k'}$, more vocabulary items are likely to be administered (especially early in the adaptive test) than would be administered using the $C_{k'}$ criterion which does not consider expected response time. Although the expected response times $t_{k'}$ may not be readily available at the item level (and may actually depend on the level of examinee ability), significant savings in test time might be obtained by using the average response time (by content area) as an approximation for $t_{k'}$.

Note that for variable length tests where each examinee is allowed to test until the time-limit has expired, the use of $\rho_{k'}$ would be expected to provide longer test-lengths and higher levels of precision than $C_{k'}$. Conversely, for fixed test-lengths (without time-limits) the use of $\rho_{k'}$ would be expected to provide shorter average completion time and lower reliabilities. Although for fixed length tests the use of $\rho_{k'}$ may lead to ability estimates with lower precision, in some situations the savings in test-time may provide adequate compensation.

Item Parameter Estimation

A number of approaches for estimating multidimensional item response functions have been proposed (Carlson, 1987; Fraser, 1988; McDonald, 1985; McKinley & Reckase, 1983; Muthén, 1984). One approach to multidimensional item parameter estimation based directly on item response theory is that of Full Information Factor Analysis (Bock, Gibbons, & Muraki, 1988). This approach specifies multidimensional estimates in which the underlying dimensions are rotated to simple structure. This rotation can be performed using either a varimax criterion (Kaiser, 1958) or the promax method (Hendrickson & White, 1964). One undesirable feature of this approach (as with all exploratory factor analytic methods) is that the dimensions of the final rotated solution may not correspond to readily interpretable factors. In addition, the computations increase exponentially with the number of factors, limiting the final solution to a maximum of five dimensions.

One alternative consists of using the unidimensional 3PL item parameter estimates to form a "multi-unidimensional" test using the strategy employed in the simulation analysis. Using this approach, each item would load on only one dimension. The loading on this dimension would correspond to the estimated discrimination parameter obtained in the unidimensional calibration. The correlation matrix $\mathbf{\Phi}$ could be estimated from the disattenuated correlations among the tests, or directly from observed responses (Mislevy, 1984). Note that approximating a multidimensional battery with a multi-unidimensional test model might provide an adequate approximation for tests in which the loadings of items on secondary dimensions are small relative to the primary loadings. This approach has been used in appropriateness measurement to form indices which provide very high rates of detection of aberrant vectors using empirical response data (Drasgow, Levine, & McLaughlin, 1991).

Although the multi-unidimensional approach is appealing in terms of its simple structure, it may suffer from at least two undesirable features. First the elements of $\mathbf{\Phi}$ may be poorly specified, depending on the method of estimation. Second, the assumption of simple structure may lead to some poorly specified loadings. However, a confirmatory approach to multidimensional item-parameter estimation may offer one compromise between the good fit to the data provided by the Full-Information approach and the simple structure obtained by the multi-unidimensional approach. A confirmatory approach, similar to the confirmatory factor analytic methods used in linear models (Bollen, 1989) would begin with a pre-specified set of zero and nonzero loadings, where the loading of each item on each dimension is based on item-content considerations. The placement of the free loadings (those allowed to differ from zero) would be made a priori, to approximate simple structure. In this way, the factor intercorrelations $\mathbf{\Phi}$ would have readily interpretable meanings and would be identified by the model without the need for a somewhat arbitrary rotation. In addition, the data could be used to alter the pattern of free and fixed discrimination parameters using a likelihood ratio test, where the fit of the model is obtained under constrained and unconstrained situations. Using this approach, items with large nonzero secondary loadings could be identified, and values for these parameters could be estimated, thus improving the fit of the model to the data. McKinley (1989) has developed a confirmatory item parameter estimation procedure in which the factors are assumed to be uncorrelated ($\mathbf{\Phi}$ is constrained to an identity matrix). This procedure, like the full-information approach suggested by Bock et al. (1988), is limited to a moderate number of dimensions. More work is necessary to identify a confirmatory approach to parameter estimation that is suitable for large item pools that span many correlated dimensions.

Exposure Control

One requirement of most adaptive tests is an item selection algorithm that limits the usage of the tests most highly informative items to avoid over exposure. The MAT and OAT simulations provided here did not incorporate exposure control algorithms.

However for operational use, an exposure control algorithm for MAT would need to be developed and implemented.

One method used to successfully control the exposure of unidimensional adaptive test items (Sympson & Hetter, 1985) can be easily adapted for use in MAT. The algorithm is based on a probabilistic approach, where each item considered for administration (using maximum information item-selection) must pass a screen. A random uniform number is generated and compared to an exposure control parameter associated with each item. If the random number is less than the exposure control parameter, then the item is administered—otherwise the item is set aside and the next most informative item is selected for consideration. This process is repeated until an item passes—then the item is administered and the response is used to update the ability after which the item selection process is repeated.

For MAT, the exposure control parameters can be calculated through a series of simulations, similar to the approach used for unidimensional adaptive tests. In computing these parameters, it is important to model examinee performance using the same item pool, item-selection, and scoring procedures to be used operationally.

Developing a Common Metric and Orientation

In an ideal item pool development effort, all items would be calibrated in a single stage. Every item would be administered to each examinee in the sample, and all items would be calibrated jointly. This approach would in principle, provide item parameters among different items which possessed the same orientation in the latent space, and which were all on a common metric.

However for multidimensional assessment, this ideal is unlikely to be achieved. When developing large pools that span several dimensions, it is often be necessary to divide the pools into small subsets of items which can be conveniently administered to examinees. Using this design however raises several issues concerning the metric and orientation of the latent dimensions. How can item parameter estimates obtained from different examinee groups be transformed to a common metric and orientation?

One appealing approach for specifying orientation is to use a confirmatory item parameter estimation procedure in which the pattern of free and fixed loadings for each item on each dimensions is specified a priori. Free loadings (nonzero parameters) would provide an indication that the item was an observed indicator of the latent dimension—a fixed zero loading would indicate otherwise. The multi-unidimensional case provides one example of this approach, where each item is allowed to load on only a single dimension.

Provided that the orientation is adequately specified, then the remaining issue for item parameter calibration involves ensuring that item parameters are placed on a common metric. Several unidimensional procedures have direct multidimensional counterparts. These procedures include: (a) random groups, (b) nonequivalent groups—joint calibration, and (c) nonequivalent groups—separate calibrations.

Random groups. If examinee groups used in the calibration are randomly equivalent, then parameters can be placed on a common metric by using the same procedure to specify the metric for each calibration group. For example, the metric can be specified by fixing the mean and variance for each dimension of the prior ability distribution. This is an extension of a method used in unidimensional marginal maximum likelihood estimation (Mislevy, 1989). By using the same specification of the prior means and variances for each calibration group, the parameters of each calibration will be placed on a common metric.

Nonequivalent groups—joint calibration. It may happen that the groups used in the calibration are not randomly equivalent. Then it is necessary for the item sets administered to each group to be linked through a series of common items. These common items provide a basis from which to perform a joint calibration of all items using all examinees. Care must be taken to ensure that the numbers of items loading on each dimension are sufficient to ensure that there is an adequate link for each dimension across all groups contributing to the calibration. Simulation studies may be required to determine the number of common items across groups necessary for linking multigroup data collection designs.

Nonequivalent groups—separate calibrations. Another option for calibrating items collected from a non-equivalent groups design involves separate marginal maximum likelihood calibrations for each group. Here the groups must be linked through a set of commonly administered items. The simplest design involves administering a common set of items to all groups. These common items would span all dimensions measured by the test. Then for one group, say Group 1, the unit and origin of each dimension can be identified though the specification of the mean and variance of the prior distribution of ability (i.e., mean vector equal to zero, and variances of each dimension equal to one). For other calibration groups, the parameters for common items are fixed at the values estimated from Group 1, and the mean vector and variances of the latent distribution are treated as parameters to be estimated along with the parameters of items unique to each calibration group.

Item Context Effects

One additional concern about the usage of MAT arises from the possibility of item context effects. For example, verbal items may function differently when they are preceded by math items than when they are preceded by other verbal items. In MAT, items from one content area can be interspersed among items of other content areas. The model presented here makes the assumption of local independence which implies that the order of presentation has no affect on the discrimination and difficulty of the item. However, items may become more or less difficult (and discriminating) depending on the content of items that precede them. Even though the mixture of item content might not affect item functioning, it may raise the level of anxiety and discomfort among examinees. Unlike typical tests, MAT may administer items of radically different content in an unpredictable sequence. On the other hand, for moderate to long tests, the mixture of item content in MAT may help break the monotony often associated with long test sessions, and help motivate the examinee. Empirical studies will be required to examine the magnitude of item context effect both in terms of its affect on examinee reactions, and its affect on MAT item selection and scoring algorithms.

Conclusions

The multidimensional Bayesian item selection and scoring procedures presented here demonstrate substantial gains in efficiency over unidimensional adaptive testing. These gains in efficiency are manifested by reduced test lengths and greater precision. In addition to increasing measurement efficiency, MAT can also be used as a tool for ensuring adequate and efficient coverage of content for examinees at different levels of proficiency. However, further study is needed before MAT can be routinely applied. This required work involves the development and refinement of item parameter estimation and exposure control methods, and the investigation of item context effects. If results are favorable, MAT may offer an attractive alternative to unidimensional CAT.

References

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

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*, 13–24.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: John Wiley & Sons.

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.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bloxom, B., & Vale, C. D. (1987, June). *Multidimensional adaptive testing: An approximate procedure for updating.* Paper presented at the meeting of the Psychometric Society, Montreal.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley & Sons.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (Research Report ONR 87-2). Iowa City, IA: The American College Testing Program.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171–191.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189–199.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373–389.

Fraser, C. (1988). *NOHARM II. A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: The University of New England, Center for Behavioral Studies.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91–115.

Hattie, J. (1981). *Decision criteria for determining unidimensionality.* Unpublished doctoral dissertation, University of Toronto, Canada.

Hendrickson, A. E., & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology, 17*, 65–70.

IMSL (1991). *International Mathematical and Statistical Libraries (Stat/Library), User's Manual,* Houston, TX: Author.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187–200.

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

McDonald, R. P. (1985). Unidimensional and multidimensional models for item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Computerized Adaptive Testing Conference* (pp. 127–148). Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Methods Program.

McKinley, R. L. (1989). *Confirmatory analysis of test structure using multidimensional item response theory* (Report No. RR-89-31). Princeton, NJ: Educational Testing Service.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods & Instrumentation, 15*, 389–390.

Miller, T., Reckase, M. D., Spray, J. A., Luecht, R., & Davey, T. (in press). *Multidimensional Item Response Theory.*

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

Mislevy, R. J. (1989). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Moreno, K. E., & Segall, D. O. (1993). CAT-ASVAB Precision. *Proceedings of the 34th Annual Conference of the Military Testing Association.* San Diego: Navy Personnel Research and Development Center.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115–132.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351–356.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude*

Battery: Development of an adaptive item pool (Technical Report 85-19). San Antonio, TX: Air Force Human Resources Laboratory.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207–230.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193–203.

Searle, S. R. (1982). *Matrix algebra useful for statistics.* New York: John Wiley & Sons.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1987). *ACAP item pools: Analysis and recommendations.* Unpublished manuscript, Navy Personnel Research and Development Center, San Diego.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item exposure rates in computerized adaptive tests.* Paper presented at the 27th Annual meeting of the Military Testing Association, San Diego, CA.

Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait.* Unpublished doctoral dissertation, Columbia University.

Wainer, H. W., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Manuscript received 5/25/94
Final version received 12/14/94