Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion

Wim J. van der Linden


Stable URL:
http://links.jstor.org/sici?sici=1076-9986%28199924%2924%3A4%3C398%3AMATWAM%3E2.0.CO%3B2-Z


Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/aera.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion

Wim J. van der Linden
University of Twente

Keywords: Adaptive Testing, Item Response Theory, Maximum-Likelihood Estimation, Multidimensionality

Adaptive testing under a multidimensional logistic response model is addressed. An algorithm is proposed that minimizes the (asymptotic) variance of the maximum-likelihood estimator of a linear combination of abilities of interest. The criterion results in a closed-form expression that is easy to evaluate. In addition, it is shown how the algorithm can be modified if the interest is in a test with a "simple ability structure". The statistical properties of the adaptive ML estimator are demonstrated for a two-dimensional item pool with several linear combinations of the abilities.

In adaptive testing, items are selected sequentially to match an updated ability estimate of the examinee. Adaptive testing algorithms for item pools calibrated under a unidimensional item response theory (IRT) model have been well investigated (e.g., Lord, 1980; Wainer, 1990), and several large-scale testing programs are in the process of introducing adaptive testing as an alternative to traditional paper-and-pencil tests. Since these programs need large item pools to guarantee measurement precision, in particular if measures to balance test content and control item exposure are implemented, violations of the assumption of unidimensionality of the item pool can be expected. Study of algorithms for adaptive testing under a multidimensional model seems to be therefore a timely matter.

The present paper is a sequel to van der Linden (1996) in which the problem of optimal assembly of a fixed test from an item pool measuring multiple abilities is addressed. The focus in this earlier paper was on an algorithm for assembling a fixed form to match a set of targets for the (asymptotic) error variance functions for the abilities subject to a large variety of constraints on the composition of the test. This paper presents an approach in which the error variance of a linear combination of the abilities is used as an item-selection criterion in tests with an adaptive format. Other approaches to multidimensional adaptive testing are presented in Fan and Hsu (April, 1996), Luecht (1996), and Segall (1996).

The author is most indebted to Wim M. M. Tielen for his computational support in preparing the numerical examples and to Cees A. W. Glas for his comments on an earlier version of this paper.
The interest in the first paper is in investigating item selection criteria based on the multivariate information matrix (Equation 9) rather than the asymptotic variance-covariance matrix (Equation 10). As will be explained later on, this paper fits the tradition in multidimensional IRT that defines a composite of interest in the ability space and then analyzes the information matrix along the direction of the composite (e.g., Ackerman, 1994, 1996; Reckase, 1985; Reckase & McKinley, 1991). In contrast, the current paper is more in line with the statistical point of view that considers the information matrix as a vehicle needed only to derive the variance-covariance matrix. In Segall (1996), the volumes of the confidence ellipsoid and the posterior credibility ellipsoid are proposed as multivariate Bayesian item selection criteria. The former is proportional to the determinant of the variance-covariance matrix. The latter is an attractive criterion because it allows for the possibility to build prior knowledge about dependencies between the ability variables into the item selection procedure. Luecht (1996) expands Segall’s approach based on the confidence ellipsoid and also offers a maximum-information criterion along a unidimensional composite defined in the ability space. We will return to the discussion of the differences between these papers and the current one in the discussion section of this paper.

The paper is organized as follows: The following section introduces the multidimensional IRT logistic model used in the presentation of the algorithm and motivates a linear combination of abilities as the parameter of interest in multidimensional adaptive testing. The subsequent section discusses the (asymptotic) variance of the estimator of a linear combination of ability parameters. Then it is proposed to minimize the variance of this estimator as a criterion for multidimensional adaptive testing, and an adaptive algorithm minimizing the variance is presented. The algorithm involves expressions defined on the item parameters that are easy to evaluate. The last section demonstrates the use of the algorithm for a two-dimensional item pool and investigates the statistical properties of the adaptive estimator for various linear combinations of the two abilities.

**Multidimensional Model**

Dichotomous response variables $U_i$ are used to denote the responses of an examinee to item $i=1, \ldots, n$. The variables take the value 1 if the response is correct and the value 0 if it is incorrect. The model is the following multivariate logistic response function:

$$p_i(\theta) = \text{Prob}\{U_i = 1 \mid \theta, a_i, d_i\} = \frac{\exp(a_i'\theta - d_i)}{1 + \exp(a_i'\theta - d_i)} \quad (1)$$

where $\theta = (\theta_1, \ldots, \theta_p, \ldots, \theta_m)$, with $-\infty < \theta_j < \infty$ for $j = 1, \ldots, m$, is a vector of $m$ ability variables, $a_i = (a_{i1}, \ldots, a_{ij}, \ldots, a_{im})$, with $a_{ij} > 0$ for $j = 1, \ldots, m$, is the vector of loadings of item $i$ on these abilities (item discriminations), and $-\infty < d_i < \infty$ is a scalar representing a linear combination of the difficulties of the item along the ability dimensions. Observe that, unlike the representation used in Reckase (1985, 1997), the item parameter is subtracted from the linear
combination of abilities rather than added. This representation allows for a
difficulty interpretation of the parameter. Use of the assumption of local inde-
pendence between responses gives the likelihood function associated with a
response vector for the model presented below. More detailed information about
the model is given in McKinley and Reckase (1983), Reckase (1985, 1997), and

It is assumed that the item parameters \( a_i \) and \( d_i \) have already been estimated,
and that the estimates are sufficiently accurate to consider them as the true
parameter values. The parameters can be estimated using the Bayesian methods
implemented in the program TESTFACT (Wilson, Wood, & Gibbons, 1984), or
through McDonald’s (1997) harmonic analysis applied to a normal approxima-
tion to the logistic function implemented in the program NOHARM (Fraser &
McDonald, 1988).

It is also assumed that the items have been calibrated on examinees showing
the same response behavior as those that take the adaptive test. Multidimension-
ality in the ability structure generally is the result of an interaction between
examinees and items and for examinees with different response behavior differ-
ent forms of multidimensionality may be obtained.

**Parameter of Interest**

It is assumed that the parameter to be estimated by the adaptive testing
procedure is a linear combination of the abilities, \( \lambda' \theta \), where \( \lambda = (\lambda_j) = (\lambda_1, \ldots, \lambda_m) \) is a vector of nonnegative weights. The components of \( \lambda \) are chosen by the
testing agent, and the CAT algorithm should be designed to measure the linear
combination defined by these components best.

The choice of this linear combination as the parameter of interest is motivated
by the following practical cases:

1. The item pool is intentionally designed to measure a predetermined
   number of abilities larger than one, which may or may not correlate. It is
   assumed that the item calibration reveals the dimensionality of the ability
   space with the desired substantive interpretation. However, the consumers
   of the test scores only want a single number to be reported. An obvious
   example is an item pool for a test to predict a future criterion of success in
   a selection problem, where the criterion is known to be multifaceted. In
   this case, the test designer will chose the weights \( \lambda_j \) to reflect the relative
   importance of the individual abilities with respect to the criterion.

2. The item pool is designed to measure only one ability but the items
   appear to be sensitive to some “nuisance abilities”. A well-know example
   is a test for mathematical ability depending on verbal abilities required to
   understand the items. This case can be dealt with by setting the weights \( \lambda_j = 0 \) for all nuisance abilities. As will become clear below, this measure
   does not neutralize the effect of the values of the nuisance parameters on
   the variance of the estimator of the intended ability directly but does
   allow for direct minimization of the latter.
3. Even though the item pool measures several abilities, different sections of the test may be required to be maximally informative with respect to different abilities, for example, because identifiable subsections of the tests are to be used for diagnosing individual abilities. The ideal here is the one of a “simple ability structure”, that is, the test is required to have different sections each measuring a different ability. As will be shown below, an adaptive test with this ability structure can be realized by choosing different values for the weights $\lambda_j$ at different stages of the procedure.

For an extended description of the above cases of multidimensional testing using the format of a fixed test, see van der Linden (1996).

*Ability Estimation*

It is assumed that $\lambda'\theta$ is estimated by the method of MLE. As is well known, for MLE it holds that:

$$\hat{\lambda'}\theta = \lambda'\hat{\theta}. \tag{2}$$

Because of local independence of the responses given $\theta$, for a response pattern $(u_1, \ldots, u_n)$ the likelihood of $\theta$ is defined as

$$L(\theta; u_1, \ldots, u_n, a_1, \ldots, a_n, d_1, \ldots, d_n) = \prod_{i=1}^{n} \text{Prob} \{ U_i = u_i | \theta, a_i, d_i \}. \tag{3}$$

The (joint) MLE of $\theta_j$, $j = 1, \ldots, m$, is the vector of values maximizing this likelihood. The likelihood equations are obtained setting the partial derivatives of the log of Equation (3) equal to zero:

$$\frac{\partial \ln L}{\partial \theta_j} = \sum_{i=1}^{n} u_i \ln p_i(\theta) + (1 - u_i) \ln (1 - p_i(\theta)) = 0, \quad j = 1, \ldots, m, \tag{4}$$

Using

$$\frac{\partial L}{\partial \theta_j} = a_{ij} p_i(\theta) [1 - p_i(\theta)], \tag{5}$$

the likelihood equations can be written as

$$\sum_{i=1}^{n} a_{ij} [u_i - p_i(\theta)] = 0, \quad j = 1, \ldots, m, \tag{6}$$

which is the common form for a model belonging to the exponential family (Andersen, 1980, sect. 3.2). The system can be solved using Newton’s method

$$\theta^{(t)} = \theta^{(t-1)} - H[\theta^{(t-1)}]^{-1} \gamma[\theta^{(t-1)}], \tag{7}$$

401
where $H(\theta^{(t-1)})$ is the Hessian of the log-likelihood function with elements

$$
\frac{\partial^2 1\ln L}{\partial \theta_g \partial \theta_h} = - \sum_{i=1}^{n} a_{ig} a_{ih} p_i(\theta)[1 - p_i(\theta)], \quad g, h = 1, \ldots, m. \tag{8}
$$

and $\gamma(\theta^{(t-1)})$ is the gradient of the log-likelihood function, that is, the vector with the first derivatives in Equation (5), both evaluated at step $t-1$. Substitution of the results from Equation (7) into Equation (2) gives the MLE of $\lambda'\theta$.

**Variance of $\hat{\lambda'}\theta$**

The asymptotic covariance matrix of the MLE of $\theta$ is given by the inverse of Fisher’s information matrix

$$
I(\theta) = - E \left( \frac{\partial^2 1\ln L(\theta; U_1, \ldots, U_n, a_1, \ldots, a_n, d_1, \ldots, d_n)}{\partial \theta_g \partial \theta_h} \right), \tag{9}
$$

with $L(\theta; U_1, \ldots, U_n, a_1, \ldots, a_n, d_1, \ldots, d_n)$ being the likelihood statistics associated with the random response vector and $\theta_p$ and $\theta_q$ any two components of $\theta$. From Equation (8), it follows that

$$
- E \left( \frac{\partial^2 1\ln L(\theta; U_1, \ldots, U_n, a_1, \ldots, a_n, d_1, \ldots, d_n)}{\partial \theta_g \partial \theta_h} \right) = \sum_{i=1}^{n} a_{ig} a_{ih} p_i(\theta)[1 - p_i(\theta)]. \tag{10}
$$

Standard techniques for matrix inversion yield the (asymptotic) covariance matrix

$$
V = \text{Var}(\hat{\theta} | \theta) = I(\theta)^{-1}, \tag{11}
$$

where the determinant of the information matrix, $|I(\theta)|$, is assumed to be not vanishing. For the linear combination $\lambda'\hat{\theta}$,

$$
\text{Var}(\lambda'\hat{\theta} | \theta) = \lambda'V\lambda. \tag{12}
$$

For a model with two ability parameters, $\theta = (\theta_1, \theta_2)$, and $\lambda = (\lambda, 1-\lambda)$, the result in Equation (12) simplifies to

$$
\begin{align*}
\text{Var}(\lambda\hat{\theta}_1 + (1 - \lambda)\hat{\theta}_2 | \theta_1, \theta_2) & \\
& = \lambda^2 \text{Var}(\hat{\theta}_1 | \theta_1, \theta_2) + (1 - \lambda)^2 \text{Var}(\hat{\theta}_2 | \theta_1, \theta_2) + 2\lambda(1 - \lambda)\text{Cov}(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2) \\
& = \left\{ \lambda^2 \sum_{i=1}^{n} a_{i1}^2 p_i(\theta_1, \theta_2)[1 - p_i(\theta_1, \theta_2)] + (1 - \lambda)^2 \sum_{i=1}^{n} a_{i2}^2 p_i(\theta_1, \theta_2)[1 - p_i(\theta_1, \theta_2)] \\
& \quad + 2\lambda(1 - \lambda) \sum_{i=1}^{n} a_{i1}a_{i2} p_i(\theta_1, \theta_2)[1 - p_i(\theta_1, \theta_2)] \right\} / |I(\theta_1, \theta_2)|. \tag{13}
\end{align*}
$$
where

\[ |I(\theta_1, \theta_2)| = \left\{ \sum_{i=1}^{n} a_{i1}^2 p_i(\theta_1, \theta_2) \left[ 1 - p_i(\theta_1, \theta_2) \right] \right\} \left\{ \sum_{i=1}^{n} a_{i2}^2 p_i(\theta_1, \theta_2) \left[ 1 - p_i(\theta_1, \theta_2) \right] \right\} \]

\[ - \left\{ \sum_{i=1}^{n} a_{i1} a_{i2} p_i(\theta_1, \theta_2) \left[ 1 - p_i(\theta_1, \theta_2) \right] \right\}^2. \]  

(14)

Note that for \( n = 2 \) the two items should not be parallel, that is, it should not hold that

\[ a_{11} = a_{21}, \]
\[ a_{12} = a_{22}, \]
\[ d_1 = d_2, \]  

(15)

because then the determinant in Equation (13) vanishes.

The ideal in adaptive testing is to have uniform measurement precision across the ability space (Lord, 1980, sect. 10.5). In the adaptive testing algorithm proposed in the next section, the asymptotic variance in Equation (12) is used to realize this ideal. A better statistical criterion of precision would be the mean-squared error (MSE) of the ability estimator, but since exact small-sample results are unknown, direct control of this criterion for shorter test lengths is impossible. However, if the item parameters are estimated precisely enough to consider them as known, the ability estimator is asymptotically unbiased and the variance and MSE of the estimator ultimately become identical measures of precision. Thus, if the test is long enough, bias vanishes and control of the (asymptotic) variance of the ability estimator suffices. The assumption of vanishing bias was met quite well for a target test length of 50 items in the empirical examples at the end of the paper.

**Adaptive Testing Algorithm**

For notational convenience, the adaptive testing algorithm is also presented for the case of two ability variables. However, the procedure immediately generalizes to linear combinations of any number of ability variables. The following definitions are needed: The items in the pool are indexed by \( i = 1, \ldots, I \). The adaptive testing procedure is assumed to be stopped after \( n \) items have been administered. Another stopping rule would be to finish the test after the value of the variance function in Equation (16) reaches a prespecified threshold. However, in practice adaptive tests are required to meet content constraints in addition to statistical criteria, and an important condition for tests to satisfy such constraints is to be of fixed length (van der Linden & Reese, 1998). The order of the items in the test is indexed by \( k = 1, \ldots, n \). Thus, \( i_k \) is the index of the item in the pool administered as the \( k \)th item in the test. Suppose \( k-1 \) items have been selected. Let \( S_k = \{ i_1, \ldots, i_{k-1} \} \) denote this set of items. Then, \( R_k = \{ 1, \ldots, I \} \setminus S_k \) is the set of items rejected so far, and item \( i_k \) has to be
selected from this set. Finally, let \( \hat{\theta}_1^k \) and \( \hat{\theta}_2^k \) be the estimators of \( \theta_1 \) and \( \theta_2 \) after \( k \) items have been administered.

The \( k \)th item is selected according to the following criterion:

\[
\min_{R_k} \left[ \text{Var}(\lambda \hat{\theta}_1^k + (1 - \lambda) \hat{\theta}_2^k | \hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \right],
\]

that is, the item is selected to minimize the variance of \( \lambda \hat{\theta}_1^k + (1 - \lambda) \hat{\theta}_2^k \) evaluated at the current estimates, which is Equation (13) with \( \theta = (\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \).

To implement the criterion, define

\[
u^j = \sum_{g=1}^{k-1} a_{i_1}^2 p_{i_1}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \left[ 1 - p_{i_1}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \right]
+ a_{i_1}^2 p_{i_1}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \left[ 1 - p_{i_1}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \right],
\]

\[
v^j = \sum_{g=1}^{k-1} a_{i_2}^2 p_{i_2}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \left[ 1 - p_{i_2}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \right]
+ a_{i_2}^2 p_{i_2}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \left[ 1 - p_{i_2}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \right],
\]

\[
w^j = \sum_{g=1}^{k-1} a_{i_3}^2 p_{i_3}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \left[ 1 - p_{i_3}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \right]
+ a_{i_3}^2 p_{i_3}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \left[ 1 - p_{i_3}(\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \right].
\]

The criterion can thus be expressed in closed form as

\[
i_k = \min_j \left\{ \left[ \lambda^2 v^j + (1 - \lambda)^2 u^j - 2\lambda(1 - \lambda)w^j \right] / \left[ u^j \nu^j - (w^j)^2 \right] ; j \in R_k \right\}.
\]

In order to select the next item, for each item in \( R_k \) one term is added to the sums in Equations (17)–(19). These terms involve both the parameters \( a_{i_1}, a_{i_2}, \) and \( d_i \) and the probability \( p_i(\theta_1, \theta_2) \), where the last quantity is evaluated at the current estimates \( \theta = (\hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}) \). The item minimizing the expression in Equation (20) is selected.

**Algorithm.** The algorithm can be summarized as follows:

1. Choose a value for \( \lambda \) reflecting the validity of the test. Preferably the value should be derived from empirical information, for example, regression weights if the test is used as a predictor in a selection problem;
2. Select item \( i_1, i_2, \) and \( i_3 \) according to some external criterion;
3. Estimate \( \theta_1 \) and \( \theta_2 \) using the MLEs from Equation (7);
4. Enter the values of the parameters of item \( i_1, i_2, \) and \( i_3 \) into Equations (17)–(19);
5. Evaluate Equations (17)–(19) for \( i \in R_k \), and select the next item to be the minimizer of Equation (20);
6. Repeat Steps 3–5 for \( k = 5, \ldots, n \).
In Step 2 the first three items are selected simultaneously; finite estimates of the ability parameters in the model in Equation (1) exist only when responses to at least three items are available. The parameter values of these items should be avoided to approximate the conditions in Equation (15) because of instability in Equation (20).

The algorithm could easily be extended to incorporate content constraints on the test (van der Linden & Reese, 1998) or to control for maximum exposure of the items (Sympon & Hetter, 1985). However, since the interest in this paper is on the statistical aspects of the algorithm, such extensions have not been pursued in the examples below.

Simple Structure. As already noted, different sections of the test may be required to have optimal measurement precision with respect to different abilities. Analogous to factor analysis, the term "simple ability structure" is used to denote this case where different sections of items in the test load onto (i.e., measure) one and only one of several dimensions. For two abilities, let $n_1$ and $n_2$ be the required numbers of items in the two sections. The proposal is to reflect this ideal by guaranteeing the test to have error variances of $\theta_1$ and $\theta_2$ proportional to the ratio of $n_1$ to $n_2$. The best way to realize this goal seems to set $\lambda$ equal to one $n_1$ times and to zero $n_2$ times while alternating item selection between the two sections from the beginning of the test. It should be noted, however, that with a multidimensional item pool the responses to each item contribute to the variance of $\theta_1$ as well as $\theta_2$, and therefore both variances must be calculated over all $n_1 + n_2$ items in the test. Therefore, either variance may become more favorable than strictly required.

Observe that this case is not equivalent to the one of choosing items from different unidimensional item pools; rather than rotating the selection of the items across unidimensional pools, the weights $\lambda_j$ are rotated across different preselected values while selecting items from a multidimensional pool.

Numerical Example

A pool of 500 items was simulated drawing random values for the parameters $a_{i1}$ and $a_{i2}$ from $U(0.0, 1.3)$ and for $d_i$ from $U(-1.3, 1.3)$. The ranges of these distributions correspond roughly to the ranges of the parameter values in a two-dimensional ACT Assessment Program Mathematics Item Pool used in van der Linden (1996) to study the performance of a linear programming model for assembling fixed-form tests. The adaptive algorithm was applied to simulated responses to a 50-item test of examinees with abilities on a two-dimensional grid defined by $\theta_1$, $\theta_2 = 2.0$, $-1.8$, $\ldots$, $2.0$. The choice of test length is slightly conservative with respect to existing CAT programs. As already mentioned, no interest existed in attempts to control the exposures rates of the items or balance test content across forms. Because the model in Equation (1) permits MLE only when at least three responses are available, the full adaptive procedure was started after responses to the first three items were simulated. The first three
items were defined to have the same parameter values \((a_{1i}, a_{2i}, d_i) = (1.2, 0.1, 0.0)\) and \((0.1, 1.2, 0.0)\) for all examinees. The log-likelihood function for the model in Equation (1) is known to have an occasional unbounded maximum. In such cases, which happened predominantly for the combination of short test lengths and extreme values of \((\theta_1, \theta_2)\), the ability estimates were truncated at \(\pm 2\). For each combination of ability values, 100 replications were produced. The study was repeated for \(\lambda = 0.250, 0.375, \) and 0.500 (larger values of \(\lambda\) were omitted because of symmetry).

Figure 1 shows the estimated bias and mean-squared error (MSE) of \(\hat{\lambda}_k^k + (1-\lambda)\hat{\theta}_2\) as a function of \(\lambda\theta_1 + (1 - \lambda)\theta_2\) for \(k = 10, 30, \) and 50 items and the different values of \(\lambda\). The bias and MSE functions are defined as \(E[(\hat{\lambda}_1 - \lambda_1) + (1 - \lambda) (\hat{\theta}_2 - \theta_2)]\) and \(E[(\lambda_1 (\hat{\lambda}_1 - \delta_1) + (1 - \lambda_2) (\hat{\theta}_2 - \delta_2))]\), respectively. These functions where estimated by averaging the (squared) estimation errors over the replications. The dominant impression from these plots is that test length is a decisive factor but that the choice of the value for \(\lambda\) hardly has any effect on the behavior of the estimator. At 10 items the estimator has an unfavorable MSE for all values of \(\lambda\theta_1 + (1 - \lambda)\theta_2\). The large MSE towards the tails of the plots is in part due to an intolerably large bias in the estimator. However, after 30 items the procedure seems to work already reasonably well. At 50 items the MSE function is uniformly low and there is hardly any bias left for all values of \(\theta\). In fact, for this test length the MSE was always smaller than .1 for all values of \(\lambda\), except for a small increase at the borders of the interval where the item difficulties probabilities are relatively underrepresented. In addition, the bias was always smaller than the MSE. As indicated earlier, these results imply that the variance function is also uniformly low and can be used as a good substitute for the MSE as criterion of measurement precision.

Alternative plots are given in Figure 2 where the bias and MSE are shown as a function of \((\hat{\theta}_1, \hat{\theta}_2)\) rather than \(\lambda_1 \theta_1 + (1 - \lambda) \theta_2\), now for \(n = 10\) and 50 and the different values of \(\lambda\). For all values of \(\lambda\), at the beginning of the test the bias function is already flat but the MSE functions show higher values for examinees that are low or high on both ability dimensions. These examinees are the ones with low probabilities of a correct response on the first items in the adaptive test; they therefore have higher values for the variances of the (ML) estimators of their abilities. This pattern is analogous to the one for the ability estimator in the unidimensional logistic models (e.g., Hambleton, Swaminathan, & Rogers, 1991, Fig. 7.4). For \(n = 50\), the two functions are both much flatter and lower, due to the fact that the adaptive algorithm has had enough time to select items with optimal properties wherever the examinee was in the ability space.

**Discussion**

As already indicated, multidimensional IRT for fixed, linear tests has a tradition of using a multivariate information measure as a tool for analyzing a
FIGURE 1. Bias and MSE of $\lambda \hat{\theta}_1^k + (1 - \lambda) \hat{\theta}_2^k$ as a function of $\lambda \hat{\theta}_1^k + (1 - \lambda) \hat{\theta}_2^k$ for: (a) $\lambda = .250$, (b) $\lambda = .375$, and (c) $\lambda = .500$: $k = 10$ (solid line); $k = 20$ (dashed line); $k = 30$ (dotted line); $k = 40$ (dashed-dotted line); $k = 50$ (bold solid line).
FIGURE 2a. Bias and MSE of $\lambda \hat{\theta}_1^k + (1 - \lambda) \hat{\theta}_2^k$ as a function of ($\theta_1$, $\theta_2$) for $\lambda = .250$ ($n = 10, 50$).

test in which the ability space is searched for directions with maximum information. These directions define the ability composite that the test is assumed to measure best. If applied to CAT, the opposite route is followed: First an ability composite is defined and then items are selected from the pool to build a test with maximum information in the direction of the composite. The current paper also defines a linear ability combination but, in contrast, the adaptive algorithm is used to minimize the variance of the ML estimator for this combination across the ability space.

The differences between the two approaches are clear when the results are plotted as a function over the ability space. In the multidimensional information approach, the plots show information measures that have a clear maximum in the desired direction (Fan & Hsu, 1996, Figs. 4-5). However, in the current paper the variance functions for the CATs become low and flat over the full ability space (Figures 2a–2c). At each point in this space the chosen linear combination of the two ability variables is measured best among all possible linear combinations.
FIGURE 2b. Bias and MSE of $\lambda \hat{\theta}_1^k + (1 - \lambda) \hat{\theta}_2^k$ as a function of ($\theta_1, \theta_2$) for $\lambda = .375$ ($n = 10, 50$).

In unidimensional IRT models, Fisher’s information in the sample is the reciprocal of the (asymptotic) variance of the ability estimator. Since the ability estimator is approximately normal with a mean converging to the true ability value, control of the elements of the information measure implies control of the sampling distribution of the estimator. In multidimensional IRT, as in any other multiparameter estimation problem, the information and variance-covariance matrix have no direct reciprocal relation. The only possibility of controlling the sampling distribution of the ability estimators is directly through the latter, as is done in Segall (1996) (determinant of variance-covariance matrix), van der Linden (1996) (minimax criterion over variance functions of the ability estimators) and in the current paper (variance function of a linear combination of abilities).

The fact that the results in the empirical example did not vary much as a function of the choice of the $\lambda$ weights implies that from a statistical point of view no preferences exist as to this choice. However, this outcome should not be
FIGURE 2c. Bias and MSE of $\lambda \hat{\theta}_1^k + (1 - \lambda) \hat{\theta}_2^k$ as a function of ($\theta_1$, $\theta_2$) for $\lambda = .500$ ($n = 10, 50$).

considered as an encouragement to ignore the role $\lambda$ when setting up an adaptive testing program. From a substantive point of view, the choice of weight is important (see the discussion of the practical cases above).

The key factor in the empirical study in this paper appears to be the length of the adaptive test. The procedure proposed seems to be practically feasible for adaptive tests of 50 items. For shorter tests, the ML estimators of $\theta_1$ and $\theta_2$ appear to be considerably biased and inefficient, even when combined linearly as in this paper. If the test length has to be short and empirical information about the correlation between the ability variables is available from external sources, it seems better to resort to a Bayesian procedure as the one in Segall (1996). This procedure allows for the possibility of building this information into the (multivariate) prior distribution for the abilities. It is well know that, provided the information leads to a location of the prior distribution at the true values of the parameters, the result is a posterior that tends to be more informative than the sampling distribution of the ML estimators. Therefore, the adaptive procedure can be expected to stabilize quicker as a function of the test length.
References


WIM J. VAN DER LINDEN is Professor of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; vanderlinden@edte.utwente.nl. He specializes in test theory, decision theory, and applied statistics.