PSYCHOMETRIKA—VOL. 74, NO. 2, 273–296
JUNE 2009
DOI: 10.1007/S11336-008-9097-5

MULTIDIMENSIONAL ADAPTIVE TESTING WITH OPTIMAL DESIGN CRITERIA FOR ITEM SELECTION

JORIS MULDER AND WIM J. VAN DER LINDEN

UNIVERSITY OF TWENTE

Several criteria from the optimal design literature are examined for use with item selection in multidimensional adaptive testing. In particular, it is examined what criteria are appropriate for adaptive testing in which all abilities are intentional, some should be considered as a nuisance, or the interest is in the testing of a composite of the abilities. Both the theoretical analyses and the studies of simulated data in this paper suggest that the criteria of A-optimality and D-optimality lead to the most accurate estimates when all abilities are intentional, with the former slightly outperforming the latter. The criterion of E-optimality showed occasional erratic behavior for this case of adaptive testing, and its use is not recommended. If some of the abilities are nuisances, application of the criterion of A_s-optimality (or D_s-optimality), which focuses on the subset of intentional abilities is recommended. For the measurement of a linear combination of abilities, the criterion of c-optimality yielded the best results. The preferences of each of these criteria for items with specific patterns of parameter values was also assessed. It was found that the criteria differed mainly in their preferences of items with different patterns of values for their discrimination parameters.

Key words: adaptive testing, Fisher information matrix, multidimensional IRT, optimal design.

1. Introduction

Unidimensional adaptive testing operates under a response model with a scalar ability parameter. Research on this type of adaptive testing has been ample. Among the topics that have been examined are the statistical aspects of ability estimation and item selection, item selection with large sets of content constraints on the test, randomized control of item exposure, removal of differential speededness, and detection of aberrant response behavior. For reviews of this research, see Chang (2004), van der Linden (2005, Chap. 9), van der Linden and Glas (2000, 2007), and Wainer (2000). Due to this research, methods of unidimensional adaptive testing are well developed, and testing organizations are now fully able to control their implementation in their testing programs.

Multidimensional item response theory (IRT) has developed gradually since its inception (e.g., McDonald, 1967, 1997; Reckase, 1985, 1997; Samejima, 1974). Its statistical tractability has been improved considerably lately, and it is now possible to use several of its models for operational testing. Although multidimensional response models are traditionally considered as a resort for applications in which unidimensional models do not show a satisfactory fit to the response data, their use has been motivated more positively recently by a renewed interest in performance-based testing and testing for diagnosis. Performance-based items typically require a range of more practical abilities. In testing for diagnosis, the goal is to extract as much information on the multiple abilities required to solve the test items as possible (e.g., Boughton, Yao, & Lewis, 2006; Yao & Boughton, 2007). Several admission and certification boards are in

The first author is now at the Department of Methodology and Statistics, Statistics, Faculty of Social Sciences, Utrecht University, Heidelberglaan 1, 3854 Utrecht, The Netherlands. The second author is now at Research Department, CTB/McGraw-Hill, Monterey, CA, USA.

Requests for reprints should be sent to Joris Mulder, Department of Research Methodology, Measurement, and Data Analysis, Twente University, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: j.mulder3@uu.nl

the process of enhancing their regular high-stakes tests with web-based diagnostic services that allow candidates to log on and get more informative diagnostic profiles of their abilities. The more informative adaptive testing format is particularly useful for this application because it is low stakes and unlikely to suffer from the security threats typical of admission and certification tests.

The first to address multidimensional adaptive testing (MAT) were Bloxom and Vale (1987), who generalized Owen's (1969, 1975) approximate procedure of Bayesian item selection to the multidimensional case. Their research did not resonate immediately with others. The only later research on multidimensional adaptive testing known to the authors is reported in Fan and Hsu (1996), Luecht (1996), Segall (1996, 2000), van der Linden (1996, 1999, Chap. 9), and Veldkamp and van der Linden (2002). Luecht and Segall based item selection for MAT on the determinant of either the information matrix evaluated at the vector of current ability estimates or the posterior covariance matrix of the abilities. In van der Linden (1999), the trace of the (asymptotic) covariance matrix of the MLEs of the abilities was minimized and the option of weighing the individual variances to control for the relative importance of the abilities was explored. The possibilities of imposing extensive sets of constraints on the item selection to deal with the content specifications of the test were examined in van der Linden (2005, Chap. 9) and Veldkamp and van der Linden (2002). The former used a criterion of minimum weighted variances for item selection; the latter the posterior expectation of a multivariate version of the Kullback–Leibler information.

Use of the determinant or trace of an information matrix or a covariance matrix as a criterion of optimality in statistical inference are standard practices in the optimal design literature, where they are known as the criteria of D-optimality and A-optimality, respectively (e.g., Silvey, 1980, p. 10). In this more general area of statistics, such criteria are used to evaluate inferences with respect to the unknown parameters in a multiple-parameter problem on a single dimension. Berger and Wong (2005) describe a variety of areas, such as medical research and educational testing, in which optimal design studies have been proven to be useful. The Fisher information matrix plays a central role in these applications because it measures the information about the unknown variables in the observations. In educational testing, for instance, the information matrix associated with a test can be optimized using the criterion of D-optimality to select a set of items from a bank with the smallest generalized variance of the ability estimators for a population of examinees. These items yield the smallest confidence region for the ability parameters. Using A-optimality instead of D-optimality yields a different selection of items because the former focuses only the variances of the ability estimators.

The answer to the question of what choice of criterion would be best is directly related to the goal of testing. As described more extensively later in this paper, different goals of MAT can be distinguished. For example, a test may be designed to measure each of its abilities accurately. But we may also be interested in a subset of the abilities and want to ignore the others. Examples of the second goal are analytic abilities in a test whose primary goal is to measure reading comprehension, or a test of knowledge of physics that appears to be sensitive to mathematical ability. In more statistical parlance, the pertinent distinction between the two cases is between the estimation of intentional and nuisance parameters. A special case of MAT with intentional ability parameters arises when the test scores have to be optimized with respect to a linear combination of them. This may happen, for instance, when the practice of having single test scores summarizing the performances on a familiar scale was established long before the use of an IRT model was introduced in the testing program. If the item domain requires a multidimensional model, it then makes sense to optimize the test scores for a linear combination of the abilities in the model with a choice of weights based on an explicit policy rather than fit a unidimensional model and accept less than satisfactory fit to the response data.

For each of these cases, a different optimal design criterion for item selection in MAT seems more appropriate. For example, as shown later in this paper, when the goal is to estimate an

intentional subset of the abilities, application of the criterion of D_s-optimality (Silvey, 1980, p. 11) to the Fisher information matrix seems leads to the best item selection. The motivation of this research was to find such matches between the different cases of MAT and the performance of optimal design criteria. In addition, to the D- and A-optimality criteria, we included a few other criteria from the optimal design literature in our research, which are less known but have some intuitive attractiveness for application in adaptive testing.

Another goal of this research was to investigate the preferences of the optimality criteria for items in the pool with specific patterns of parameter values. The results should help to answer such questions as: Will the criterion for selection in a MAT program with nuisance abilities select only items that are informative about the intentional abilities? Or are there any circumstances in which they also select items that are mainly sensitive to a nuisance ability? Understanding the preferences of item-selection criteria for different patterns of parameter values is important for the assembly of optimal item pools for the different cases of MAT when there exists a choice of items. In principle, such information could help us to prevent overexposure and underexposure of the items in the pool and reduce the need of using more conventional measures of item-exposure control (Sympson & Hetter, 1985; van der Linden & Veldkamp, 2007).

Finally, we report some features of the Fisher information matrix and its use in adaptive testing that have been hardly noticed hitherto and also illustrate the use of the criteria empirically using simulated response data.

2. Response Model

The response model used in this paper is the multidimensional 3-parameter logistic (3PL) model for dichotomously scored responses. The model gives the probability of a correct response to item i by an examinee with p-dimensional ability vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ as

$$\begin{aligned} P_i(\boldsymbol{\theta}) &\equiv P\left(U_i = 1 | \boldsymbol{\theta}, \mathbf{a}_i, b_i, c_i\right) \\ &\equiv c_i + \frac{1 - c_i}{1 + \exp(-\mathbf{a}_i \cdot \boldsymbol{\theta} + b_i)}, \end{aligned} \tag{1}$$

where $\mathbf{a}_i$ is a vector with the item-discrimination parameters corresponding to the abilities in $\boldsymbol{\theta}$, b_i is a scalar representing the difficulty of item i, and c_i is known as the guessing parameter of item i (i.e., the height of the lower asymptote of the response function). Note that b_i is not a difficulty parameter in the same sense as a unidimensional IRT model; in the current parameterization, it is a function of both the difficulties and the discriminating power of the item along each of the ability dimensions. Further, note that due to rotational indeterminacy of the $\boldsymbol{\theta}$-space, the components of $\boldsymbol{\theta}$ do not automatically represent the desired psychological constructs. However, such issues are dealt with when the item pool is calibrated, and we can assume that a meaningful orientation of the ability space has been chosen. Finally, note that (1) is just a model for the probability of a correct answer by a fixed test taker. Particularly, it is not used as part of a hierarchical model in which $\boldsymbol{\theta}$ is a vector of random effects. Therefore, the model should not be taken to imply anything with respect to a possible correlation structure between the abilities in some population of test takers; for instance, it does not force us to decide between what are known as orthogonal and oblique factor structures in factor analysis.

The vector of discrimination parameters, $\mathbf{a}_i$, can be interpreted as the relative importance of each ability for answering the item correctly. As is often done, we assume that the item parameters have been estimated with enough precision to treat them as known. The probability of an incorrect response will be denoted as $Q_i(\boldsymbol{\theta}) = 1 - P_i(\boldsymbol{\theta})$. Because this model is not yet identifiable, additional restrictions are necessary that fix the scale, origin, and orientation of $\boldsymbol{\theta}$. In practical

applications of IRT models, testing organizations maintain a standard parameterization through the use of parameter linking techniques that carry the restrictions from the calibration of one generation of test items to the next. For more details about the model, see Reckase (1985, 1997) and Samejima (1974). Other multidimensional response models are available in the literature, but the model in (1) is a direct generalization of the most popular unidimensional logistic model in the testing industry. In addition, its choice allows us to compare our results with those in a key reference on item selection in MAT as Segall (1996).

The following additional notation will be used throughout this article:

N: size of the item pool;
n: length of the adaptive test;
$l = 1, \ldots, p$: components of ability vector $\boldsymbol{\theta}$;
$i = 1, \ldots, N$: items in the pool;
$k = 1, \ldots, n$: items in the test;
i_k: item in the pool administered as the kth item in the test;
S_{k-1}: set of first $k-1$ administered items;
R_k: $\{1, \ldots, N\} \backslash S_{k-1}$, i.e., set of remaining items in the pool.

For a vector $\mathbf{u}_{k-1}$ of responses on the first $k-1$ items, the maximum likelihood estimate (MLE) of the ability, denoted by $\hat{\boldsymbol{\theta}}^{k-1}$, is defined as

$$\hat{\boldsymbol{\theta}}^{k-1} \equiv \arg\max_{\boldsymbol{\theta}} f(\mathbf{u}_{k-1}|\boldsymbol{\theta}), \tag{2}$$

where

$$f(\mathbf{u}_{k-1}|\boldsymbol{\theta}) = \prod_{j=1}^{k-1} P_{i_j}^{u_{i_j}}(\boldsymbol{\theta}) Q_{i_j}^{1-u_{i_j}}(\boldsymbol{\theta}), \tag{3}$$

is the likelihood function with the item responses modeled conditionally independent given $\boldsymbol{\theta}$. The MLE can found by setting the derivative of the logarithm of (3) equal to zero and solve the system for $\boldsymbol{\theta}$ using a numerical method such as Newton–Raphson (e.g., Segall, 1996) or an EM algorithm (Tanner, 1993, Chap. 4). The likelihood function may not have a maximum (e.g., when only correct or incorrect item responses are observed), or a local instead of a global maximum may be found. Such problems are rare for adaptive tests of typical length, though.

3. Fisher Information

The Fisher information matrix is a convenient measure of the information in the observable response variables on the vector of ability parameters $\boldsymbol{\theta}$. For item i, the matrix is defined as

$$\begin{aligned} \mathbf{I}_i(\boldsymbol{\theta}) &\equiv -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(U_i|\boldsymbol{\theta})\right] \\ &= \frac{Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i]^2}{P_i(\boldsymbol{\theta})(1-c_i)^2}\mathbf{a}_i\mathbf{a}_i^T, \end{aligned} \tag{4}$$

with $\mathbf{a}_i^T$ the transpose of the (column) vector of discrimination parameters. This expression reveals some interesting features of the item information matrix:

- The item information matrix depends on the ability parameters only through the response function $P_i(\boldsymbol{\theta})$.

- The matrix has rank one.
- Each element in the matrix has a common factor, which will be denoted as

$$g(\boldsymbol{\theta}; \mathbf{a}_i, b_i, c_i) = \frac{Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i]^2}{P_i(\boldsymbol{\theta})(1 - c_i)^2}. \tag{5}$$

  This function of $\boldsymbol{\theta}$ will be discussed in the following section.
- The sum of the elements of the matrix is equal to

$$g(\boldsymbol{\theta}; \mathbf{a}_i, b_i, c_i)\left(\sum_{l=1}^{p} a_{il}\right)^2.$$

  This equality shows the important role played by the sum of the discrimination parameters in the total amount of information in the response to an item.

The information matrix of a set of S items is equal to the sum of the item information matrices, i.e.,

$$\mathbf{I}_S(\boldsymbol{\theta}) = \sum_{i \in S} \mathbf{I}_i(\boldsymbol{\theta}). \tag{6}$$

The additivity follows from the conditional independence of the responses given $\boldsymbol{\theta}$ already used in (3). Although the item information matrix $\mathbf{I}_i(\boldsymbol{\theta})$ of each item in S has rank 1, the rank of $\mathbf{I}_S(\boldsymbol{\theta})$ is equal to p (unless the items in S have the same proportional relationship between the discrimination parameters).

The use of the information matrix is mainly motivated by the large-sample behavior of the MLE of $\boldsymbol{\theta}$, which is known to be distributed asymptotically as

$$\hat{\boldsymbol{\theta}} \sim N\big(\boldsymbol{\theta}_0, \mathbf{I}_S^{-1}(\boldsymbol{\theta}_0)\big) \tag{7}$$

with $\boldsymbol{\theta}_0$ the true ability and $\mathbf{I}_S^{-1}(\boldsymbol{\theta}_0)$ the inverse of the information matrix evaluated at $\boldsymbol{\theta}_0$. More generally, it holds for the covariance matrix $\Sigma(\boldsymbol{\theta}_0)$ of any unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ that $\Sigma(\boldsymbol{\theta}_0) - \mathbf{I}_S^{-1}(\boldsymbol{\theta}_0)$ is positive semi-definite. The inverse information matrix can thus be considered as the multivariate generalization of the Cramér–Rao lower bound on the variance of estimators (Lehmann, 1999, Section 7.6).

In test theory, it is customary to consider $\mathbf{I}_S(\boldsymbol{\theta})$ and $\mathbf{I}_i(\boldsymbol{\theta})$ as functions of $\boldsymbol{\theta}$ and refer to them as the test and item information matrix, respectively. By substituting $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in (6), an estimate of these matrices is obtained. When evaluating the selection of the kth item in the adaptive test using (6), the amount of information about $\boldsymbol{\theta}$ can be expressed as the sum of the test information matrix for the $k-1$ items already administered and the matrix for candidate item i_k,

$$\mathbf{I}_{S_{k-1}}\big(\hat{\boldsymbol{\theta}}^{k-1}\big) + \mathbf{I}_{i_k}\big(\hat{\boldsymbol{\theta}}^{k-1}\big). \tag{8}$$

Criteria of optimal item selection should thus be applied to (8). For example, Segall (1996) proposed to select the item that maximizes the determinant of (8). This candidate gives the largest decrement in volume of the confidence ellipsoid of the MLE $\hat{\boldsymbol{\theta}}^{k-1}$ after $k-1$ observed responses. As already noted, a maximum determinant of an information matrix is known as D-optimality in the optimal design literature. Before dealing with such criteria in more detail, we take a closer look at the item information matrix.

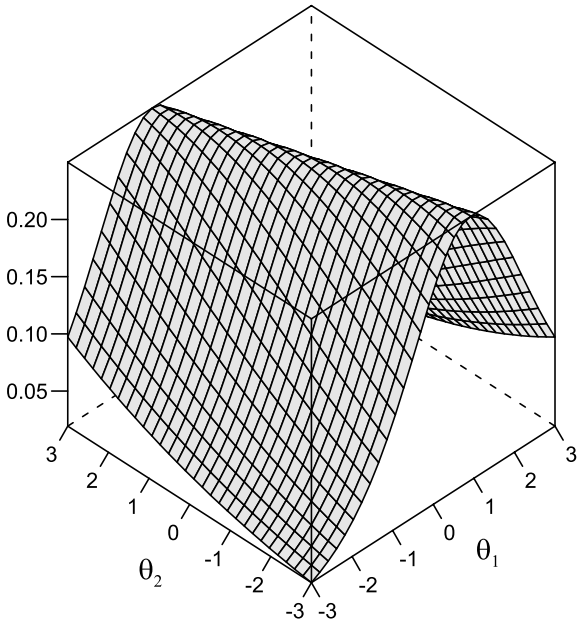

FIGURE 1.
Surface of $g(\boldsymbol{\theta}, \mathbf{a}, d, c)$ with $\mathbf{a} = (1, 0.3)$, $d = 0$, and $c = 0$. (Note: $\tilde{g}$ is a cross-section of g perpendicular to $a_1\theta + a_2\theta = 0$.)

3.1. Item Information Matrix in Multidimensional IRT

The item information matrix in (4) can be written as

$$\mathbf{I}_i(\boldsymbol{\theta}) = g(\boldsymbol{\theta}; \mathbf{a}_i, b_i, c_i) \begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \dots & a_{i1}a_{ip} \\ a_{i1}a_{i2} & a_{i2}^2 & \dots & a_{i2}a_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1}a_{ip} & a_{i2}a_{ip} & \dots & a_{ip}^2 \end{bmatrix}, \tag{9}$$

with function g given in (5). Thus, the information matrix consists of two factors: (i) function g and (ii) matrix $\mathbf{a}_i\mathbf{a}_i^T$ with elements based on the discrimination parameters.

The focus of the next sections is on the comparison of different optimality criteria for item selection in different cases of MAT. Each of these criteria maps the item information matrix onto a one-dimensional scale. Because $g(\boldsymbol{\theta}; \mathbf{a}_i, b_i, c_i)$ is a common factor in all elements of the information matrix, the criteria basically differ in how they deal with the second factor $\mathbf{a}_i\mathbf{a}_i^T$. On the other hand, g is function of $\boldsymbol{\theta}$, and it is instructive to analyze its shape, which is done in this section. As an example, in Figure 1, g is plotted for an item with parameters $\mathbf{a}_i = (1, 0.3)$, $b_i = 0$, and $c_i = 0$ over a two-dimensional ability space $\boldsymbol{\theta}$.

Observe that g depends on $\boldsymbol{\theta}$ only through the response function $P_i(\boldsymbol{\theta})$ in (1). Because $P_i(\boldsymbol{\theta})$ is constant when $\boldsymbol{\theta} \cdot \mathbf{a}_i$ is, the same applies to g. For example, for the item displayed in Figure 1, the values of g do not depend on the abilities as long as $\theta_1 + 0.3\theta_2$ is constant. This feature can be used to reparameterize $g(\boldsymbol{\theta}; \mathbf{a}_i, b_i, c_i)$ into a one-dimensional function $\tilde{g}(\theta; \mathbf{a}_i, b_i, c_i)$ with a new θ perpendicular to $\boldsymbol{\theta} \cdot \mathbf{a}_i = 0$. As shown in Figure 1, $\tilde{g}$ is just a cross-section of g.

The new function $\tilde{g}$ is obtained by substituting

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_p) = \left(\frac{a_{i1}}{\|\mathbf{a}_i\|}\theta, \dots, \frac{a_{ip}}{\|\mathbf{a}_i\|}\theta \right)$$

into the response function in (1), which results in

$$\tilde{P}_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp(-\|\mathbf{a}_i\|\theta + b_i)}, \tag{10}$$

where $\|\mathbf{a}_i\| = \sqrt{a_{i1}^2 + \cdots + a_{ip}^2}$ is the Euclidean norm of $\mathbf{a}_i$. Thus, the reparameterization leads to a new unidimensional response model which differs from the regular unidimensional 3PL model only in the replacement of its discrimination parameter by the Euclidean norm of the vector with the discrimination parameters from the multidimensional model.

By definition, the maximizer of $\tilde{g}$, denoted here as $\theta_{\max}$, is the ability value for which the item provides most information. It can be determined by solving $\frac{\partial}{\partial\theta}\tilde{g}(\theta; \mathbf{a}_i, b_i, c_i) = 0$ for θ. The result is

$$\theta_{\max} = \begin{cases} \frac{b_i - \log(\frac{-1+\sqrt{1+8c_i}}{4c_i})}{\|\mathbf{a}_i\|} & \text{for } c_i > 0, \\ \frac{b_i}{\|\mathbf{a}_i\|} & \text{for } c_i = 0. \end{cases} \tag{11}$$

In addition,

$$\tilde{g}(\theta_{\max}; \mathbf{a}_i, b_i, c_i) = \begin{cases} \frac{16c_i(1-c_i)(-1+\sqrt{1+8c_i})}{(3+\sqrt{1+8c_i})(4c_i-1+\sqrt{1+8c_i})^2} & \text{for } c_i > 0, \\ 0.25 & \text{for } c_i = 0. \end{cases} \tag{12}$$

These results enable us to use the intuitions developed for the unidimensional 3PL model for the current multidimensional generalization of it. First, from (11), it follows that $\theta_{\max}$ increases with the guessing parameter of the item. Hence, when an item has a higher chance of guessing the correct answer, it should be used for more able test takers. Second, the difficulty parameter serves as a location parameter for the item in the direction perpendicular to $\boldsymbol{\theta} \cdot \mathbf{a}_i = 0$. The parameter is scaled by the Euclidean norm of the discrimination parameters of the item. Third, (12) shows that the maximum value of $\tilde{g}(\theta; \mathbf{a}_i, b_i, c_i)$, and hence also of $g(\boldsymbol{\theta}; \mathbf{a}_i, b_i, c_i)$, only depends on the guessing parameter. Fourth, $g(\theta_{\max}; \mathbf{a}_i, b_i, c_i)$ decreases with increasing c_i. This can be shown by calculating the derivative of $\tilde{g}(\theta_{\max}; \mathbf{a}_i, b_i, c_i)$ with respect to c_i, which is omitted here. Consequently, the maximum values of the elements in the item information matrix depend only on the discrimination parameters through the matrix $\mathbf{a}_i\mathbf{a}_i^T$. This conclusion reconfirms the critical role of this matrix as the second factor of (9).

4. Item Selection Criteria for MAT

When testing multiple abilities, different cases of multidimensional testing should be distinguished (van der Linden, 2005, 1999, Section 8.1). This article focuses on three cases; the others can be considered as minor variations of them:

1. All abilities in the ability space are intentional. The goal of the test is to obtain the most accurate estimates for all abilities.
2. Some abilities are intentional and the others are nuisances. This case arises, for instance, when a test of knowledge of physics has items that also require language skills, but the goal of this test is not to estimate any language skill.
3. All abilities measured by the test are intentional, but the interest is only in a specific linear combination of them. As already explained, this case occurs in practice when the test is multidimensional, but for historic reasons, the examinees' performances are to be reported in the form of single scores.

Different optimal design criteria based on the item information matrix rank the same set of items differently for test takers of equal ability. The choice of criterion for item selection in adaptive testing should therefore be in agreement with the goal of the MAT program. However, it is not immediately clear which criterion is best for each of the above cases of MAT. For the first case, both D-optimality and A-optimality seem reasonable choices. The former seeks to minimize the generalized variance, the latter the sum of the variances of the ability estimators. But it is unclear how they will behave in the two other cases of MAT. Both criteria will be analyzed in more detail below. In addition, the usefulness of the less-known criterion of E-optimality, D_s-optimality, and c-optimality will be investigated. In order to obtain expressions that are relatively easy to interpret, the criteria are derived for a three-dimensional ability space. The conclusions for ability spaces of higher dimensionality are similar. Wherever possible, for notational simplicity, the argument $\hat{\boldsymbol{\theta}}^{k-1}$ in the information matrices in (8) is omitted.

4.1. All Abilities Intentional

The goal is to obtain accurate estimates of the separate abilities in $\boldsymbol{\theta}$. For this case, the following three optimality criteria are likely candidates.

4.1.1. D-Optimality This criterion maximizes the determinant of (8). Hence, it selects the kth item to be

$$\arg\max_{i_k \in R_k} \det(\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k}). \tag{13}$$

Using the factorization in (9), the criterion can be expressed as

$$\begin{aligned}\arg\max_{i_k \in R_k} g\big(\hat{\boldsymbol{\theta}}^{k-1}; \mathbf{a}_{i_k}, b_{i_k}, c_{i_k}\big)\big(&a_{i_k1}^2 \det(\mathbf{I}_{S_{k-1}[1,1]}) + a_{i_k2}^2 \det(\mathbf{I}_{S_{k-1}[2,2]}) + a_{i_k3}^2 \det(\mathbf{I}_{S_{k-1}[3,3]}) \\ &- 2a_{i_k1}a_{i_k2} \det(\mathbf{I}_{S_{k-1}[1,2]}) - 2a_{i_k2}a_{i_k3} \det(\mathbf{I}_{S_{k-1}[2,3]}) - 2a_{i_k1}a_{i_k3} \det(\mathbf{I}_{S_{k-1}[1,3]})\big),\end{aligned} \tag{14}$$

where $\mathbf{I}_{S_{k-1}[l_1,l_2]}$ is the submatrix of $\mathbf{I}_{S_{k-1}}$ when omitting row l_1 and column l_2.

In matrix algebra, $\mathbf{I}_{S_{k-1}[l_1,l_2]}$ is referred to as a cofactor and its determinant is known as the minor. Observe that the square of the discrimination parameter corresponding to θ_1 is multiplied by $\det(\mathbf{I}_{S_{k-1}[1,1]})$, which can be interpreted as the current amount of information about the two other abilities, θ_2 and θ_3. Similar relationships hold for a_{i_k2} and a_{i_k3}. Consequently, the criterion tends to select items with a large discrimination parameter for the ability with a relatively large (asymptotic) variance for its current estimator. The criterion thus has a built-in "minimax mechanism": it tends to pick the items that minimize the variance of the estimator lagging behind most. The same behavior has been observed for D-optimal item calibration designs (Berger & Wong, 2005, p. 15). As a result, the difference between the sampling variances of the estimators for the two abilities tend to be negligible toward the end of the test. This is precisely what we may want when both abilities are intended to be measured.

From (14), it can also be concluded that items with large discrimination parameters for more than one ability are generally not informative. Consequently, the criterion of D-optimality tends to prefer items that are sensitive to a single ability over items sensitive to multiple abilities.

Segall (1996, 2000) proposed using a Bayesian version of D-optimality for MAT that evaluates the determinant of the posterior covariance matrix at the posterior modes of the abilities (instead of the determinant of the information matrix at the MLEs). Assuming a multivariate normal posterior, he showed the result to be

$$\arg\max_{i_k \in R_k} \det\big(\mathbf{I}_{S_{k-1}}(\tilde{\boldsymbol{\theta}}^{k-1}) + \mathbf{I}_{i_k}(\tilde{\boldsymbol{\theta}}^{k-1}) + \boldsymbol{\Sigma}_0^{-1}\big), \tag{15}$$

where $\boldsymbol{\Sigma}_0$ is the prior covariance matrix of $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}^{k-1}$ is the posterior mode after $k-1$ items have been administered.

4.1.2. A-Optimality This criterion seeks to minimize the sum of the (asymptotic) sampling variances of the MLEs of the abilities, which is equivalent to selecting the item that minimizes the trace of the inverse of the information matrix:

$$\arg\min_{i_k \in R_k} \operatorname{trace}\left((\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k})^{-1}\right) = \arg\max_{i_k \in R_k} \frac{\det(\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k})}{\sum_{l=1}^{3} \det([\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k}]_{[l,l]})}. \tag{16}$$

A-optimality results in an item-selection criterion that contains the determinant of the information matrix as an important factor. Its behavior should thus be largely similar to that of D-optimality.

Analogous to Segall's (1996, 2000) proposal, a Bayesian version of A-optimality could be formulated by adding the inverse of a prior covariance matrix to (16) and evaluating the result at a Bayesian point estimate of $\boldsymbol{\theta}$ instead of the MLE. But this option is not pursued here any further.

4.1.3. E-Optimality The criterion of E-optimality maximizes the smallest eigenvalue of the information matrix, or equivalently, the generalized variance of the ability estimators along their largest dimension. The criterion has gained some popularity in the literature on optimal regression design; for an application to optimal temperature input in microbiological studies, where the criterion has shown to work efficiently, see Bernaerts, Servaes, Kooyman, Versyck, & Van Impe (2002). A disadvantage of the criterion might be its lack of robustness in applications with sparse data. Due to its complexity, the expression of the smallest eigenvalue of the matrix $\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k}$ is omitted here.

In spite of the popularity of the criterion in other applications, it may behave unfavorably when used for item selection in MAT. As shown in Appendix A, the contribution of an item with equal discrimination parameters to the test information vanishes when the sampling variances of the ability estimators have become equal to each other. This fact contradicts the fundamental rule that the average sampling variance of an MLE should always decrease after a new observation. Using E-optimality for item selection in MAT may therefore result in occasionally bad item selection, and its use is not recommended. The simulation studies later in this paper confirm this conclusion.

4.1.4. Graphical Example In order to get a first impression of the behavior of the three optimality criteria, their surfaces for two 2-dimensional items are plotted. As indicated earlier, the discrimination parameters play a crucial role. We therefore ignore possible differences between the difficulty and guessing parameters and consider the following two items:

$$\text{Item 1:} \quad a_1 = (0.5, 1), b_1 = 0, \text{ and } c_1 = 0, \tag{17}$$
$$\text{Item 2:} \quad a_2 = (0.8, 0), b_2 = 0, \text{ and } c_2 = 0. \tag{18}$$

Item 1 is sensitive to both abilities but for Item 2 the second ability does not play any role in answering the item correctly. The current information matrix is fixed at

$$\mathbf{I}_{S_{k-1}} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix} \tag{19}$$

for all ability values. This choice enable us to clearly see the difference between the surfaces.

The surfaces for the three criteria are shown in the left-hand panels of Figure 2, whereas the right-hand panels display some of their contours as a function of the discrimination parameters

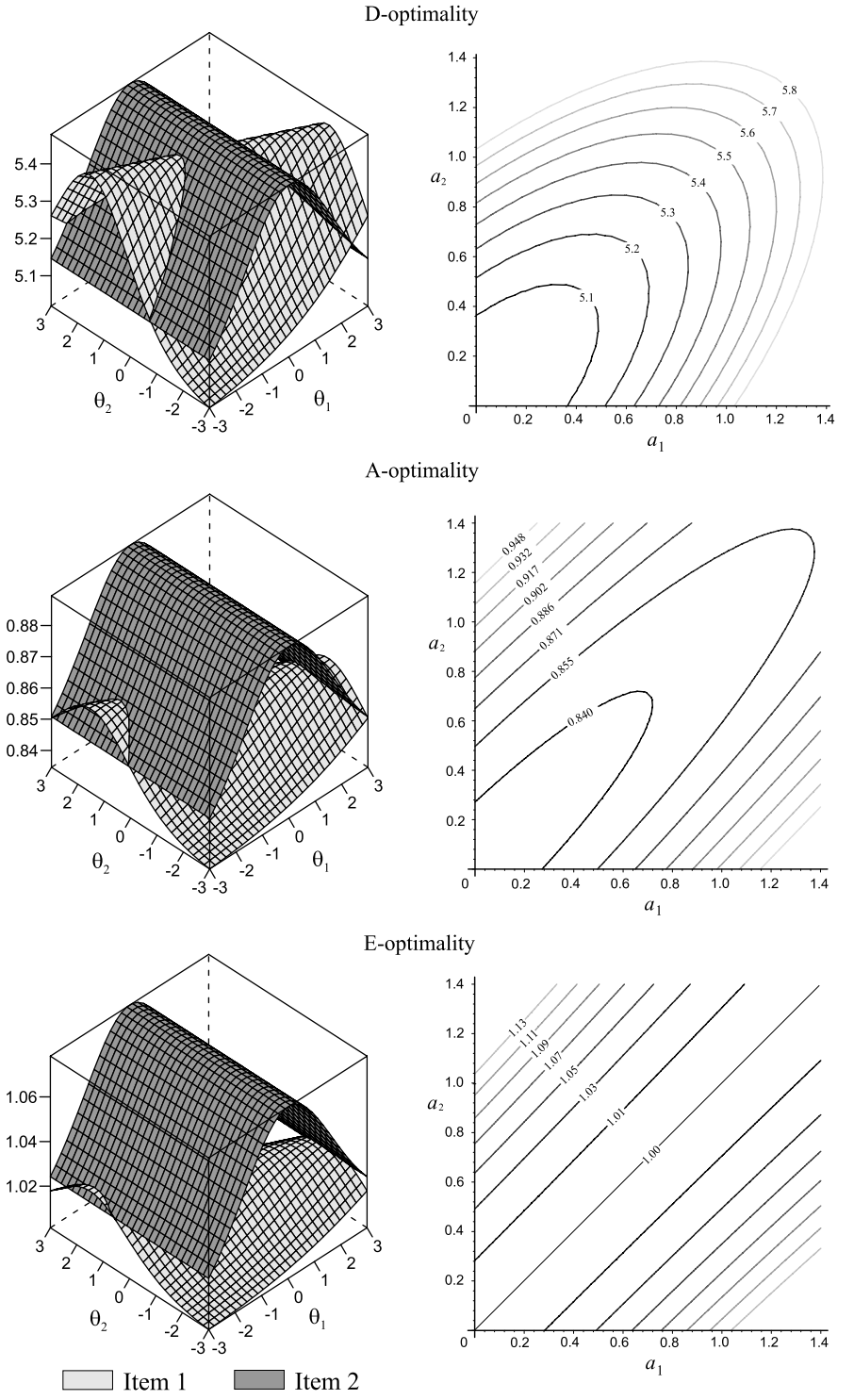

FIGURE 2.
Surfaces of the criteria of D-, A-, and E-optimality for Item 1 and Item 2 (*left-hand panels*) and contours of the same criteria as a function of the discrimination parameters (a_1, a_2) for a person with average ability $\boldsymbol{\theta} = \mathbf{0}$ (*right-hand panels*). (Note: $b = 0$ and $c = 0$.)

for abilities fixed at $\boldsymbol{\theta} = \mathbf{0}$. (Note that for A-optimality, we plotted the argument of the right-hand side of (16), so that a higher surface is equivalent to a more informative item.) The shapes of the surfaces seem to be entirely determined by the common factor $g(\boldsymbol{\theta}; \mathbf{a}_i, b_i, c_i)$ of the elements of the item information matrix. The differences between the three criteria are caused only by the different ways in which they map $\mathbf{a}_i \mathbf{a}_i^T$ onto a one-dimensional space. Each criterion finds Item 2, which tests a single ability, most informative. But the preference for this item is strongest for E-optimality and weakest for D-optimality. This conclusion follows comparing an item with discrimination parameters $(a_1, 0)$ with one that has (a, a) but the same item information score, for instance, $\mathbf{a} = (0.5, 0)$ and $\mathbf{a} = (0.64, 0.64)$ (D-optimality) and $\mathbf{a} = (0.5, 0)$ and $\mathbf{a} = (1.36, 1.36)$ (A-optimality).

For the more extreme values of $\boldsymbol{\theta} = (2, -2)$ and $(2, 2)$, the contours in Figure 3 show some surprising shapes. For instance, if $a_2 = 0$, an increase of a_1 does not always result in an increase of the criterion. Thus, for items that do not show any discrimination with respect to one of the abilities, the occurrence of extreme values of the MLEs of θ_1 and θ_2 in the beginning of an adaptive test is likely to result in inappropriate item selection for the criteria of D- and A-optimality. Obviously, such items should not be admitted to the pool. Also, the independence of the criterion of E-optimality of the discrimination parameters when they are equal $(a_1 = a_2)$ is demonstrated by its contours. As already indicated, this behavior of the criterion of E-optimality does not meet our intuitive idea of information in an item.

4.2. Nuisance Abilities

When the first s abilities of the ability vector $\boldsymbol{\theta}$ are intentional and the last $p - s$ abilities are nuisances, D_s-optimality (Silvey, 1980, p. 11) seems to reflect the goal of this case of MAT. In this case, our interest goes to the vector $\mathbf{A}^T\boldsymbol{\theta}$ with $\mathbf{A}^T = [I_s \ 0]$, where I_s is a $s \times s$ identity matrix. D_s-optimality selects the item

$$\arg\max_{i_k \in R_k} \det\left(\mathbf{A}^T (\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k})^{-1} \mathbf{A}\right)^{-1}. \tag{20}$$

Instead of maximizing the determinant of $(\mathbf{A}^T (\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k})^{-1} \mathbf{A})^{-1}$, the trace of its inverse could be minimized. The criterion would then be called A_s-optimality. Below we consider two instances of this case for a three-dimensional ability vector $\boldsymbol{\theta}$.

4.2.1. θ_1 and θ_2 Intentional and θ_3 a Nuisance Ability Let θ_1 and θ_2 be intentional abilities and θ_3 a nuisance ability. Hence,

$$\mathbf{A}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The criterion in (20) can then be expressed as

$$\arg\max_{i_k \in R_k} \left(\det\left(\left[\left(\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k}\right)^{-1}\right]_{[3,3]}\right)\right)^{-1}. \tag{21}$$

Note that θ_3 is not ignored in (21) because the criterion is based on the inverse of the information matrix $\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k}$ instead of the matrix itself. As a result of taking the determinant, items that mainly test a single ability are generally most informative.

However, the criterion does not always select items that only discriminate highly with respect to one of the intentional abilities. This point is elaborated in Appendix B, where we show that when the amount of information about the intentional abilities is high relative to the amount of information about all abilities, i.e., $\det((\mathbf{I}_{S_{k-1}})_{[3,3]}) > \det(\mathbf{I}_{S_{k-1}})$, the criterion reveals a tendency to select items that discriminate highly with respect to the nuisance ability. Under these

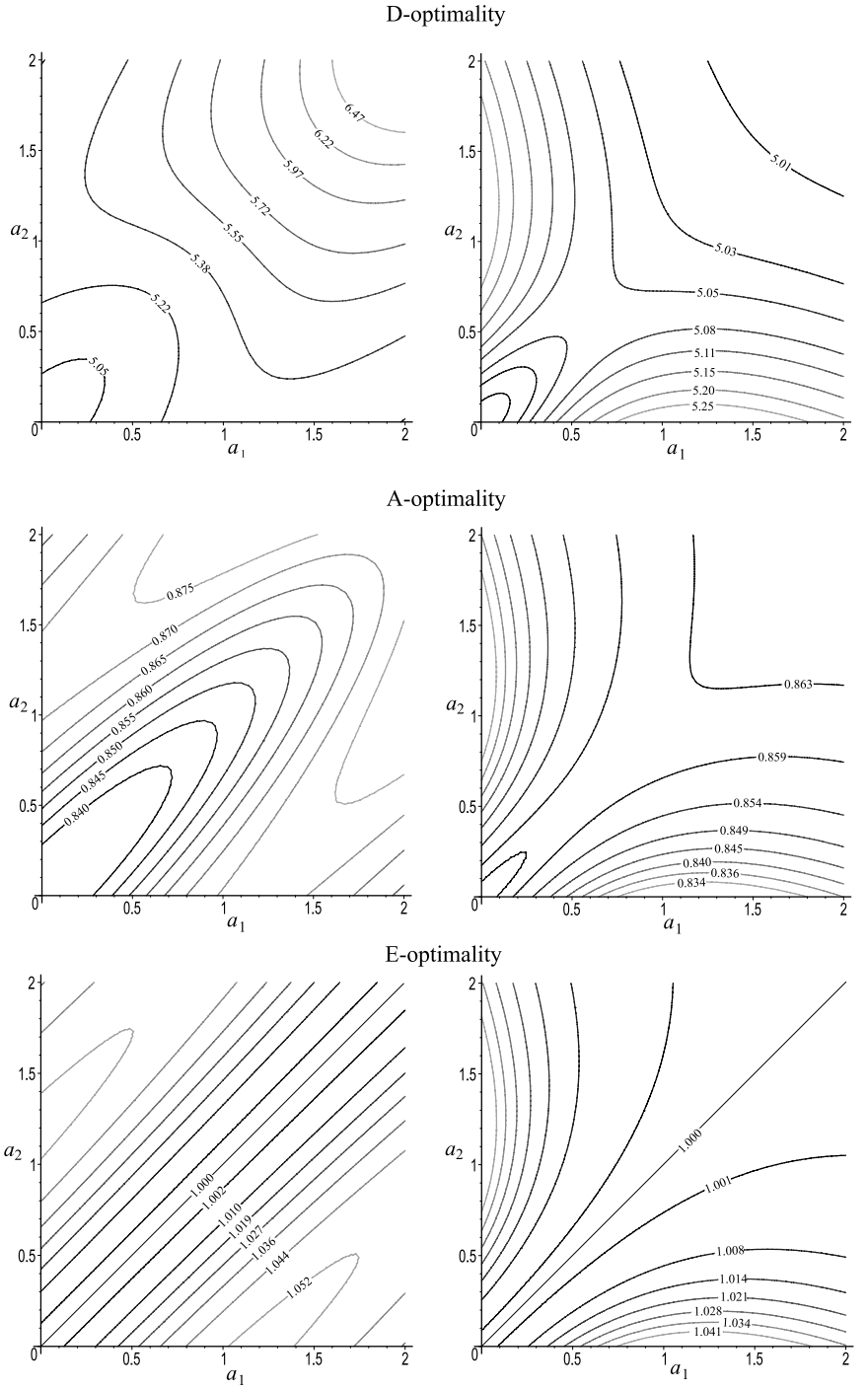

FIGURE 3.
Contours of the criteria of D-, A-, and E-optimality as a function of the discrimination parameters of the item for $\boldsymbol{\theta} = (-2, 2)$ (*left-hand panels*) and $\boldsymbol{\theta} = (2, 2)$ (*right-hand panels*). (Note: $b = 0$ and $c = 0$.)

conditions, the sampling variance of the estimator of the nuisance abilities is relatively large, and the selection of such items results in the largest decrease of the generalized variance of the intentional abilities. This type of behavior was also observed in a study with simulated data reported later in this paper. Similarly, it can be shown that the behavior of the criterion of A_s-optimality is similar to that of D_s-optimality.

4.2.2. θ_1 Intentional and θ_2 and θ_3 Nuisance Abilities Because θ_1 is the only intentional ability, $\mathbf{A}^T = [1\ 0\ 0]$. Consequently, (20) selects the item that minimizes the sampling variance of θ_1, that is,

$$\arg\min_{i_k \in R_k} \left[(\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k})^{-1} \right]_{(1,1)}, \tag{22}$$

where $[\,\cdot\,]_{(1,1)}$ denotes element $(1, 1)$ of the matrix. In Appendix B, we show that this criterion generally selects items that highly discriminate with respect to the intentional ability, θ_1, except when the amount of information about the nuisance abilities is relatively low, i.e., $\det((\mathbf{I}_{S_{k-1}})_{[1,1]})$ is small. In this case, (22) prefers selecting an item that highly discriminates with respect to the nuisance abilities. Similar behavior was observed for the case of two intentional and one nuisance ability.

Observe that the criterion of A_s-optimality selects the same items as (22).

4.3. Composite Ability

This case of MAT occurs when the items in the item pool measure multiple abilities but only an estimate of a specific linear combination of the abilities,

$$\theta_c = \boldsymbol{\lambda} \cdot \boldsymbol{\theta} = \sum_{l=1}^{p} \lambda_l \theta_l, \tag{23}$$

is required, with $\boldsymbol{\lambda}$ a vector of (nonnegative) weights for the importance of the separate abilities. In order to maintain a standardized scale for θ_c, often $\sum_{l=1}^{p} \lambda_l = 1$ is used.

Because the response probability in (1) depends on the abilities only through the linear combination $\mathbf{a}_i \cdot \boldsymbol{\theta}$, an item is informative for a composite θ_c when $\mathbf{a}_i = \alpha_i \boldsymbol{\lambda}$, for some large constant $\alpha_i > 0$. This claim is immediately clear when comparing the multidimensional model after substituting $\mathbf{a}_i = \alpha_i \boldsymbol{\lambda}$ in (1) with the unidimensional IRT model,

$$\begin{aligned} P_i(\boldsymbol{\theta}) &\equiv c_i + \frac{1 - c_i}{1 + \exp(-\alpha_i \boldsymbol{\lambda} \cdot \boldsymbol{\theta} + b_i)} \\ &= c_i + \frac{1 - c_i}{1 + \exp(-\alpha_i \theta_c + b_i)}, \end{aligned}$$

where α_i would the discrimination parameter of the unidimensional IRT model.

According to Silvey (1980, p. 11), c-optimality applies when we wish to obtain an accurate estimate of a linear combination of unknown parameters. For the current application to MAT, the criterion can be shown to be equal to

$$\arg\min_{i_k \in R_k} \boldsymbol{\lambda}^T \left(\mathbf{I}_{S_{k-1}}\left(\hat{\boldsymbol{\theta}}^{k-1}\right) + \mathbf{I}_{i_k}\left(\hat{\boldsymbol{\theta}}^{k-1}\right) \right)^{-1} \boldsymbol{\lambda} = \arg\max_{i_k \in R_k} \left(\boldsymbol{\lambda}^T \left(\mathbf{I}_{S_{k-1}}\left(\hat{\boldsymbol{\theta}}^{k-1}\right) + \mathbf{I}_{i_k}\left(\hat{\boldsymbol{\theta}}^{k-1}\right) \right)^{-1} \boldsymbol{\lambda} \right)^{-1}. \tag{24}$$

Indeed, this criterion prefers items with discrimination parameters that reflect the weights of importance in the composite ability, i.e., $\mathbf{a}_i \propto \boldsymbol{\lambda}$. The preference is demonstrated for a two-dimensional ability vector with equal weights $\lambda_1 = \lambda_2 = 1$ in Figure 4. (Note that we plotted

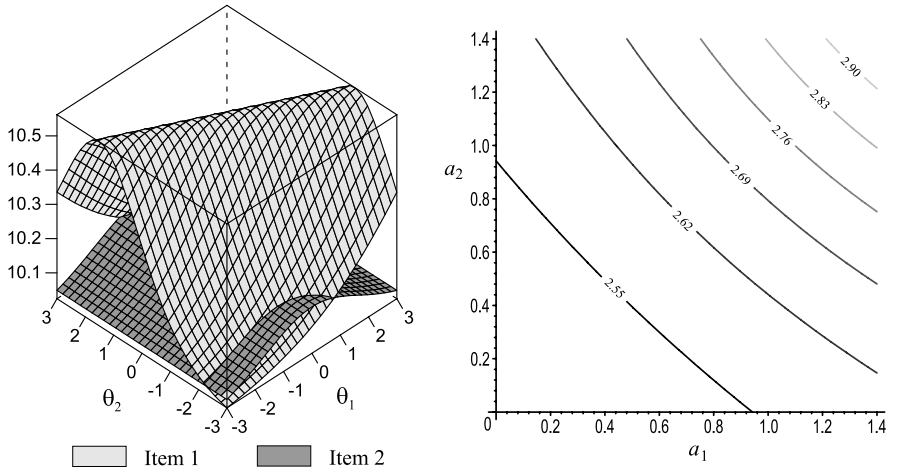

FIGURE 4.
Surfaces of the criterion of c-optimality for Items 1 and 2 (*left-hand panels*) and contours of the criterion of c-optimality for the same items for $\boldsymbol{\theta} = \mathbf{0}$ (*right-hand panels*). (Note: $b = 0$ and $c = 0$.)

the argument in the right-hand side of (24), so that a larger outcome can be interpreted as a more informative item.) Item 1 is generally more informative because $\boldsymbol{\lambda} \cdot \mathbf{a}_1$ is larger than $\boldsymbol{\lambda} \cdot \mathbf{a}_2$. Furthermore, unlike the criteria of D-, A-, and E-optimality, which yielded concave contours (see Figure 2), the contours in Figure 4 are convex. Thus, for this criterion, an item that tests several abilities simultaneously with $\mathbf{a}_i \propto \boldsymbol{\lambda}$ is generally more informative than an item with a preference for a single ability.

5. Simulation Study

In order to assess the influence of the item-selection criteria on the accuracy of the ability estimates, we conducted a study with simulated data. Each of the three cases of MAT discussed in the previous section were simulated to see whether the proposed selection criteria resulted in the best results. Also, we were interested in seeing if the proposed selection rule for adaptive testing with a nuisance ability would result in more accurate estimation of the intentional ability than when all abilities are to be considered as intentional. Similar interest existed in the estimation of a specific linear combination of the abilities using the criterion in (24).

The second goal of this study was to find out what combinations of discrimination parameters for an item were generally most informative for each of the three cases of adaptive testing. This was done by counting how often each item was administered for each selection criterion. The information is helpful when designing an item pool for a given case of multidimensional adaptive testing and item-selection criterion.

Finally, we were interested in seeing whether each of the optimality criterion resulted in the best value of the specific quantity it optimizes, for instance, whether the determinant of the covariance matrix (that is, the generalized variance) at the end of the test was actually smallest for the criterion of D-optimality.

All simulations were done for the case of a two-dimensional vector $\boldsymbol{\theta}$. Analysis of the higher-dimensional expressions for the optimality criteria in the previous sections shows that the dimensionality of the ability space is reflected only in the order of the information matrices and not in the structure of the expression. For instance, it is easy for the reader to verify that the argument that revealed the peculiar behavior of the criterion of D_s-optimality in Appendix B for the case

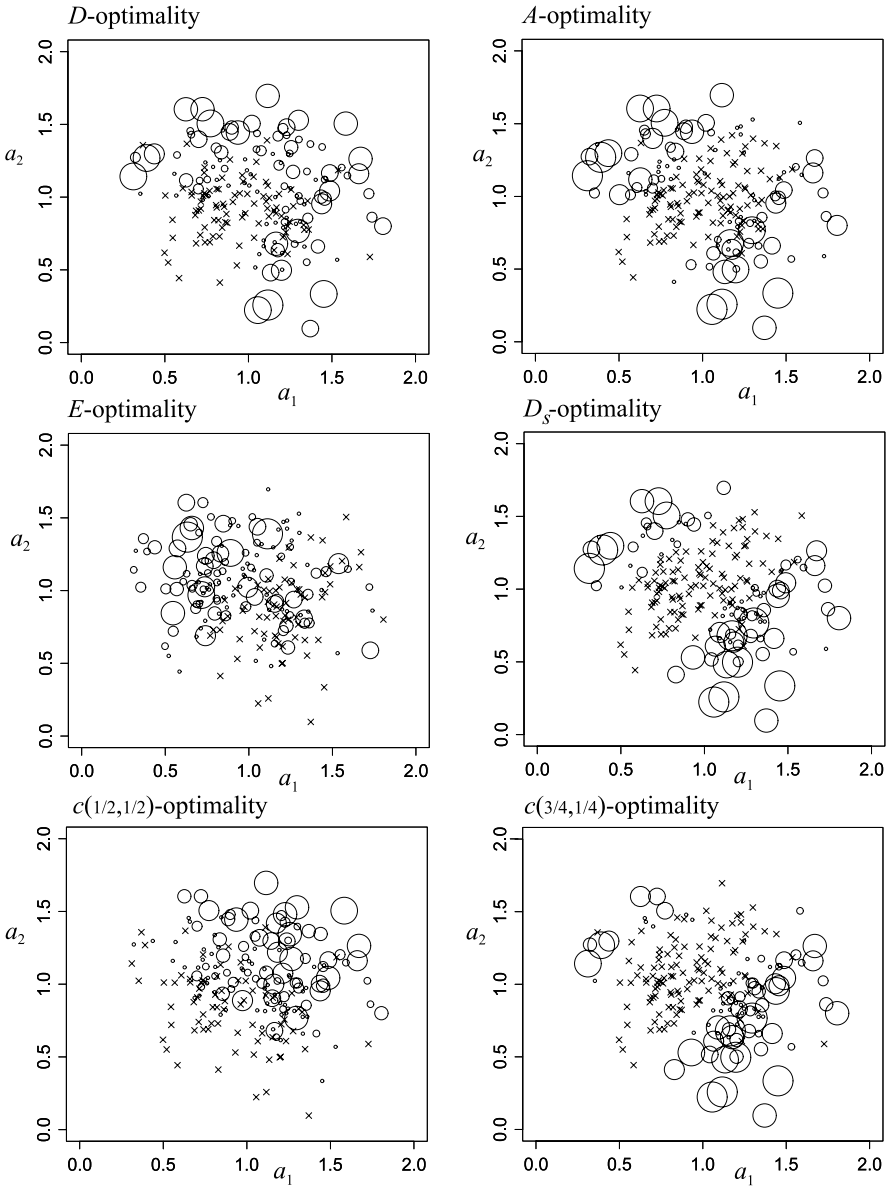

FIGURE 5.
Empirical frequencies of the discrimination parameters of the items selected for the different optimality criteria. (Note: the size of the circles is proportional to the frequency of selection; × denotes items that were never selected.)

of a three-dimensional space with intentional and nuisance abilities, which forced us reject this criterion, holds equally well for a two-dimensional space with one intentional and one nuisance ability. We therefore assume the behavior of the item-selection criteria to be similar for multidimensional adaptive testing with any number of dimensions.

5.1. Design of the Study

The behavior of these criteria is further illustrated by the empirical frequencies of item selection plotted against the item discrimination parameters of the items in Figure 5. In these graphs,

TABLE 1.
First five items administered for the simulated adaptive tests.

	Item 1	Item 2	Item 3	Item 4	Item 5
a_{i1}	2	2	0	0	2
a_{i2}	0	0	2	2	2
b_i	5	−5	5	−5	0
c_i	0	0	0	0	0

each symbol represents an item with its discrimination parameters as coordinates. The size of the circle is proportional to the number of times the item was selected for administration. Items that where never selected are symbolized as "×." The preference of items with a high discrimination parameter for a single ability is stronger for A- than for D-optimality. The difference might explain why the former slightly outperformed the latter in terms of accuracy of ability estimation in this case. The frequencies of the difficulty and guessing parameters are omitted because they were as expected for A-optimality and D-optimality: For both criteria, the distributions of the difficulty parameters were close to uniform. Also, smaller guessing parameters were selected much more frequently than larger parameters. We also prepared the same plots as in Figure 5 for the conditional distributions of the frequencies of item selection given the abilities in this study. But since they were generally similar to the marginal distributions, they are omitted here.

The item pool consisted of 200 items that were generated according to $a_1 \sim N(1, 0.3)$, $a_2 \sim N(1, 0.3)$, $b \sim N(0, 3)$, and $10c \sim \text{Bin}(3, 0.5)$. None of the items had negative discrimination parameters. The MLEs of the abilities were calculated using a Newton–Raphson algorithm. To ensure the existence of the MLEs, each adaptive test began with the five fixed items displayed in Table 1. When a MLE was obtained, 30 items were selected adaptively from the pool using the item-selection rules described in this article. For each combination of $\theta_1 = -1, 0, 1$ and $\theta_2 = -1, 0, 1$, a total of 100 adaptive test administrations were simulated. Hence, a total of 900 tests were administered for each selection rule. The final MLEs of the abilities of interest were compared with the test takers' true abilities by calculating their average bias

$$\text{Bias}(\hat{\theta}_l) = \frac{1}{100}\sum_{j=1}^{100}(\hat{\theta}_{j,l} - \theta_l)$$

and mean squared error (MSE)

$$\text{MSE}(\hat{\theta}_l) = \frac{1}{100}\sum_{j=1}^{100}(\hat{\theta}_{j,l} - \theta_l)^2,$$

where $\hat{\theta}_{j,l}$ is the final estimate of ability $l = 1, 2$ of the jth simulated test taker and θ_l is this test taker's true value for ability l.

In order to have a baseline for our comparisons, we also simulated test administrations in which the adaptive selection of the 30 items was replaced by random selection from the pool. Table 2 shows the estimated bias and Table 3 shows the MSE for these administrations (columns labeled "R"). The MSEs reveal close to uniform precision for the estimation of θ_1 and θ_2 across their range of values, with $(-1, -1)$ as an exception. This finding indicates that the range of difficulty of the items in the pool was wide enough to cover the abilities in this study. As a consequence of the specific combinations of the randomly generated item parameters, apparently the items in the pool tended to be more effective for estimating θ_2 than θ_1.

TABLE 2.
Bias of the estimates of θ_1 and θ_2 for different item selection rules: Random selection (Column R), D-optimality (Column D), A-optimality (Column A), E-optimality (Column E), and D_s-optimality (Column D_s) with θ_1 intentional and θ_2 a nuisance ability.

θ_1	θ_2	Bias($\hat{\theta}_1$)					Bias($\hat{\theta}_2$)				
		R	D	A	E	D_s	R	D	A	E	D_s
1	1	0.108	−0.101	−0.005	0.169	−0.009	−0.074	0.056	0.016	0.148	0.025
1	0	−0.158	−0.048	−0.097	−0.052	−0.021	0.270	0.097	0.185	0.550	0.070
1	−1	−0.181	−0.199	−0.123	−0.755	−0.082	0.272	0.137	0.113	0.467	0.145
0	1	0.095	0.040	0.244	0.898	0.117	0.004	−0.045	−0.174	−0.180	−0.158
0	0	−0.153	0.054	0.108	−0.048	0.005	0.277	−0.050	−0.099	0.073	0.090
0	−1	−0.168	−0.012	−0.028	−0.919	−0.061	0.226	0.033	0.072	0.327	0.063
−1	1	0.416	0.138	0.189	0.989	0.006	−0.451	−0.202	−0.182	−0.484	−0.003
−1	0	0.264	0.073	−0.002	−0.007	0.102	−0.256	−0.030	−0.002	−0.190	−0.083
−1	−1	−0.138	0.180	0.052	−0.236	0.005	0.168	−0.113	0.018	0.010	0.130
Average		0.009	0.014	0.038	0.003	0.007	0.048	−0.013	−0.006	0.080	0.031

TABLE 3.
MSE of the estimates of θ_1 and θ_2 for different item selection rules: Random selection (Column R), D-optimality (Column D), A-optimality (Column A), E-optimality (Column E), and D_s-optimality (Column D_s) with θ_1 intentional and θ_2 a nuisance ability.

θ_1	θ_2	MSE($\hat{\theta}_1$)					MSE($\hat{\theta}_2$)				
		R	D	A	E	D_s	R	D	A	E	D_s
1	1	0.686	0.407	0.481	0.248	0.360	0.694	0.399	0.464	0.414	0.472
1	0	0.800	0.417	0.482	0.797	0.425	0.863	0.516	0.456	1.305	0.584
1	−1	0.926	0.509	0.399	2.337	0.348	0.972	0.482	0.351	1.263	0.429
0	1	0.909	0.380	0.503	1.549	0.451	0.819	0.417	0.462	0.777	0.557
0	0	1.107	0.502	0.504	1.625	0.419	0.969	0.454	0.543	0.924	0.617
0	−1	1.128	0.409	0.416	1.640	0.282	1.045	0.388	0.391	0.941	0.501
−1	1	1.006	0.447	0.417	2.582	0.350	0.997	0.535	0.395	1.063	0.526
−1	0	0.941	0.429	0.427	0.589	0.451	0.764	0.484	0.474	0.820	0.537
−1	−1	0.710	0.392	0.326	0.092	0.361	0.686	0.412	0.330	0.517	0.507
Average		0.911	0.432	0.439	1.273	0.383	0.868	0.454	0.429	0.892	0.525

5.2. θ_1 and θ_2 Intentional

Our analysis of the behavior of the selection criteria for this case of MAT suggested using either D- or A-optimality. For completeness, we also included the less favorable E-optimality criterion in the study. The results for these criteria are given in Tables 2 and 3 (columns labeled "D," "A," "E"). As Table 3 shows, the criteria of D- and A-optimality resulted in substantial improvement of ability estimation over random selection. Furthermore, using the criterion of A-optimality resulted on average in slightly more accurate ability estimation than that of D-optimality. As expected, the criterion of E-optimality performed badly (even worse than the baseline). In fact, this finding definitively disqualifies E-optimality as a criterion for item selection in adaptive testing.

The poor results for E-optimality as an item-selection criterion are explained by the inappropriate behavior of the criterion described in Appendix B. For examinees of extreme ability, the criterion tended to select items of opposite difficulty; hence, its low efficiency.

5.3. θ_1 Intentional and θ_2 a Nuisance

For this case, the criterion of D_s-optimality selects items minimizing the asymptotic variance of the intentional ability θ_1. Tables 2 and 3 (columns labeled "D_s") display the results from our study. The MSE for the estimator of θ_1 was much more favorable than for the criteria of A- and D-optimality in the preceding section (θ_1 and θ_2 both intentional). As expected, these results were obtained at a much larger MSE for the estimator of θ_2. This finding points at the fact that the presence of a second intentional ability introduces a trade-off between their two estimators, and consequently to less favorable behavior for either of them.

The preferences of the current criterion for the pairs of discrimination parameters for the items was as expected (see Figure 5). The majority of the items selected mainly tested the intentional ability; only few had a preference for the nuisance ability. Generally, the preferences for the difficulty and guessing parameters by this criterion appeared to be very similar to those for the earlier criteria of D-optimality and A-optimality.

5.4. Composite Ability

Two different composite ability combinations were addressed. For the first composite ability, the criterion of c-optimality with weights $(\frac{1}{2}, \frac{1}{2})$ was used as selection rule, i.e., $\theta_c = \frac{1}{2}\theta_1 + \frac{1}{2}\theta_2$. The bias and MSE of the estimator of this composite ability are given in Table 4. For comparison, we also calculated the bias and MSE of a plug-in estimator with substitution of the MLEs of θ_1 and θ_2 from the earlier simulations with D- and A-optimality into the linear composite, the reason being a similar interpretation of these criteria in terms of weights of importance of θ_1 and θ_2. On average, c-optimality with weights $(\frac{1}{2}, \frac{1}{2})$ yielded the highest accuracy for the estimates of θ_c. Of course, all results for the estimators of the composite were obtained at the price of a larger MSE for the estimators of the separate abilities. (The latter are not shown here; their averages were 0.649 and 0.616 for the estimators of θ_1 and θ_2, respectively.)

Second, a composite ability with unequal weights was considered: $\theta_c = \frac{3}{4}\theta_1 + \frac{1}{4}\theta_2$. In this composite, the first ability is considered to be more importance than the second. Again, the items were selected using c-optimality with weights $(\frac{3}{4}, \frac{1}{4})$ as criterion. The bias and MSE of the estimator are given in Table 5, which also shows the results for the plug-in estimator with the MLEs of θ_1 and θ_2 from the earlier simulations with D_s- and c-optimality with weights $(\frac{1}{2}, \frac{1}{2})$. Note that D_s-optimality is equivalent to c-optimality with weights $(1, 0)$. Table 5 shows that

TABLE 4.
Bias and MSE of the estimate of $\theta_c = \frac{1}{2}\theta_1 + \frac{1}{2}\theta_2$ for adaptive testing with D-, A-, and c-optimality with weights $(\frac{1}{2}, \frac{1}{2})$ as item selection criterion.

θ_1	θ_2	Bias($\hat{\theta}_c$)			MSE($\hat{\theta}_c$)		
		D	A	$c(\frac{1}{2}, \frac{1}{2})$	D	A	$c(\frac{1}{2}, \frac{1}{2})$
1	1	−0.022	0.006	0.024	0.034	0.037	0.038
1	0	0.025	0.044	0.024	0.030	0.057	0.032
1	−1	−0.031	−0.005	0.002	0.037	0.046	0.029
0	1	−0.002	0.035	−0.009	0.034	0.051	0.029
0	0	0.002	0.005	−0.014	0.034	0.062	0.029
0	−1	0.010	0.022	0.009	0.043	0.059	0.043
−1	1	−0.032	0.004	0.008	0.030	0.037	0.026
−1	0	0.022	−0.002	0.025	0.038	0.039	0.028
−1	−1	0.034	0.035	−0.023	0.042	0.046	0.043
Average		0.001	0.016	0.005	0.036	0.048	0.033

c-optimality with the weights $(\frac{3}{4}, \frac{1}{4})$ resulted in the smallest average MSE for this composite ability.

Figure 5 also displays the empirical frequencies of the discrimination parameters selected in these simulations with the composite abilities. (The distributions for the difficulty and guessing parameters are omitted because they were similar to those for the previous cases.) The criterion of c-optimality with the weights $(\frac{1}{2}, \frac{1}{2})$ had strong preference for items with a large value for $a_1 + a_2$. This finding reflects the fact that we tested the simple sum $\theta_1 + \theta_2$. Consequently, when the item was sensitive to θ_1 or θ_2 only, it tended to be ignored by the criterion. For the case of c-optimality with the weights $(\frac{3}{4}, \frac{1}{4})$, the distribution of the discrimination parameters was similar to that for D_s-optimality. This result makes sense because the weights were now closer to the case of $(1, 0)$ implied by the use of D_s-optimality. It also explains the small difference between the MSEs for D_s-optimality and c-optimality with the weights $(\frac{3}{4}, \frac{1}{4})$ in Table 5.

5.5. *Average Values of Optimality Criteria*

Table 6 shows the average determinant, trace, largest eigenvalue, first diagonal element, the weighted sum with $\boldsymbol{\lambda}_1 = (\frac{1}{2}, \frac{1}{2})^T$, and the weighted sum with $\boldsymbol{\lambda}_2 = (\frac{3}{4}, \frac{1}{4})^T$ of the final covariance matrix $\mathbf{I}_{S_n}^{-1}(\hat{\boldsymbol{\theta}})$ at the end of the simulated adaptive tests for each of the selection criteria. Except for E-optimality, each of the criteria produced the smallest average value for the specific quan-

TABLE 5.
Bias and MSE of the estimate of $\theta_c = \frac{3}{4}\theta_1 + \frac{1}{4}\theta_2$ for adaptive testing with D_s-optimality, c-optimality with weights $(\frac{3}{4}, \frac{1}{4})$, and c-optimality with weights $(\frac{1}{2}, \frac{1}{2})$ as item selection criterion. Note that D_s-optimality is equivalent with c-optimality with weights $(1, 0)$.

θ_1	θ_2	Bias($\hat{\theta}_c$)			MSE($\hat{\theta}_c$)		
		D_s	$c(\frac{3}{4}, \frac{1}{4})$	$c(\frac{1}{2}, \frac{1}{2})$	D_s	$c(\frac{3}{4}, \frac{1}{4})$	$c(\frac{1}{2}, \frac{1}{2})$
1	1	−0.001	−0.026	0.024	0.122	0.125	0.164
1	0	0.002	−0.058	0.024	0.125	0.133	0.207
1	−1	−0.025	−0.090	0.002	0.110	0.136	0.211
0	1	0.048	0.029	−0.009	0.127	0.114	0.207
0	0	0.026	−0.019	−0.014	0.126	0.142	0.193
0	−1	−0.030	0.000	0.009	0.081	0.104	0.171
−1	1	0.004	0.088	0.008	0.105	0.091	0.198
−1	0	0.056	0.052	0.025	0.147	0.103	0.210
−1	−1	0.037	−0.001	−0.023	0.118	0.108	0.161
Average		0.013	−0.003	0.005	0.118	0.117	0.191

TABLE 6.
Average value of the quantities optimized by each of the selection criteria at the end of the adaptive tests.

	D-opt	A-opt	E-opt	D_s-opt	$c(\boldsymbol{\lambda}_1)$-opt	$c(\boldsymbol{\lambda}_2)$-opt
$\det(\mathbf{I}^{-1})$	0.069*	0.081	2.46	0.089	0.110	0.112
$\text{trace}(\mathbf{I}^{-1})$	1.007	0.944*	4.32	1.007	1.779	1.249
$\max\{\text{eigenvalues}(\mathbf{I}^{-1})\}$	0.933	0.849*	3.673	0.909	1.713	1.153
$(\mathbf{I}^{-1})_{(1,1)}$	0.499	0.469	2.76	0.439*	0.891	0.470
$\boldsymbol{\lambda}_1^T \mathbf{I}^{-1} \boldsymbol{\lambda}_1$	0.037	0.048	0.437	0.052	0.033*	0.060
$\boldsymbol{\lambda}_2^T \mathbf{I}^{-1} \boldsymbol{\lambda}_2$	0.152	0.152	1.168	0.133	0.248	0.124*

*Smallest element in a row; $\boldsymbol{\lambda}_1 = (\frac{1}{2}, \frac{1}{2})^T$; $\boldsymbol{\lambda}_2 = (\frac{3}{4}, \frac{1}{4})^T$.

tity optimized by it. For instance, the criterion of D-optimality resulted in the smallest average determinant of the final covariance matrix (= smallest generalized variance) among all criteria.

6. Conclusions

Both our theoretical analyses and the results from the study with simulated data allow us to draw the following conclusions:

1. When all abilities are intentional, the criterion of A-optimality tends to result in the most accurate MLEs for the separate abilities. But the results for D-optimality were close. The most informative items measure mainly one ability, i.e., have one large discrimination parameter and small parameters for the other abilities. Furthermore, both criteria tend to "minimax": when the estimator of one of the abilities has a small sampling variance, they develop a preference for items that are highly informative about the other abilities. Consequently, the accuracy of the final estimates of a sufficiently long tests are approximately equal.
2. When one of the abilities is of interest and the others should be considered as a nuisance, item selection based on D_s-optimality (or A_s-optimality) seems to result in the most accurate estimates for the intentional ability. The accuracy of the estimator of the intentional ability tends to be higher than when all abilities are considered as intentional. The advantage is obtained at the price of less accurate estimation of the nuisance ability. (But, of course, this is something we should be willing to pay.) Again, items that measure only the intentional ability are generally most informative. But when the current inaccuracy of the estimator of a nuisance ability becomes too large relative to that of the intentional abilities, the dependency of the latter on the former becomes manifest and an occasional preference for an item mainly sensitive to the nuisance ability emerges.
3. When the goal is to estimate a linear combination of the abilities, c-optimality with weights $\boldsymbol{\lambda}$ proportional to the coefficients in the composite ability results in the most accurate MLE of the combination. The criterion has a preference for items when the proportion of the discrimination parameters reflects the weights in the combination.

All conclusions were based on analyses of criteria for a three-dimensional abilities. As already indicated, generalization to higher dimensionality does not involve any new obstacle. However, it should be observed that these conclusions only hold for an item pool that allows free selection from all possible combinations of item parameters (as in our simulation study). For instance, when a two-dimensional item pool would consists only of items with a small discrimination parameter for θ_1 and a large parameter for θ_2, but only the former is intentional, the MSE of its estimator might be larger than that of the nuisance ability, even when our suggestions for the choice of criterion are followed.

In fact, even when the item pool has no constraints, we often are forced to impose constraints on the item selection that may have the same effect. One obvious example is when we need to constrain the item selection to guarantee that the content specifications for the test are satisfied (van der Linden, 2005, Chap. 9) and the content attributes of the items correlate with their statistical parameters. Another example is when item selection is constrained to deal with potential overexposure of some of the items in the pool, for instance, when using the Sympson–Hetter (1985) exposure control method or the selection is constrained more directly using item-ineligibility constraints (van der Linden & Veldkamp, 2007). Because the item-exposure rates are typically correlated with the discrimination parameters of the items, exposure control is expected to have an even stronger impact on our conclusions.

As a next step, it would be interesting to investigate item selection more closely when using other information measures than Fisher's. The most likely candidate is Kullback–Leibler information (Chang & Ying, 1996; Veldkamp & van der Linden, 2002). For example, it would be

interesting to see if an application of the criterion of A-optimality would then also prefer items that mainly test a single ability and c-optimality would prefer items with large sums of discrimination parameters. A confirmation of the findings in this article for other information measures would made them more robust. However, we do not expect the criterion of E-optimality criterion to show improved behavior for other measures. Both theoretically and empirically, we found its behavior to be too erratic to warrant application in real-world adaptive testing.

Appendix A: E-optimality

The anomaly is illustrated for a test information matrix that after $k-1$ items has become equal to

$$\mathbf{I}_{S_{k-1}} = \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & d \end{bmatrix}.$$

This matrix occurs when each of the $k-1$ items tested a single ability and the sampling variances of their estimators have become equal.

Now, consider a candidate item with equal discrimination parameters for each ability, that is,

$$\mathbf{I}_{i_k} = g\big(\hat{\boldsymbol{\theta}}^{k-1}; \mathbf{a}, b, c\big) \begin{bmatrix} a^2 & a^2 & a^2 \\ a^2 & a^2 & a^2 \\ a^2 & a^2 & a^2 \end{bmatrix}.$$

The eigenvalues of $\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k}$ are then equal to

$$\boldsymbol{\lambda} = \big(d + 3g\big(\hat{\boldsymbol{\theta}}^{k-1}; \mathbf{a}, b, c\big)a^2, d, d\big) \quad \Rightarrow \quad \min_{l=1,2,3}\{\lambda_l\} = d.$$

So, the selection of the new item would not lead to any change of the current information matrix $I_{S_{k-1}}$ (that is, $I_{S_{k-1}} = I_{S_{k-1}} + I_{i_k}$). According to the criterion of E-optimality, the response to the candidate item contains no information about the ability parameters, which contradicts with the fundamental idea that the average (asymptotic) sampling variance of MLEs strictly decreases with the size of the sample.

Appendix B: Nuisance Ability

We explore the behavior of the criterion in (20) for the case of both θ_1 and θ_2 intentional and θ_3 a nuisance ability as well as that of θ_1 intentional and the other two a nuisance ability. The versions of the criterion for these two cases are denoted by $D_s^{(12)}$ and $D_s^{(1)}$, respectively.

B.1 θ_1 and θ_2 Intentional and θ_3 a Nuisance Ability

Candidate item i_k is selected to maximize

$$D_s^{(12)}(\mathbf{a}_{i_k}, b_{i_k}, c_{i_k}) = \Big(\det\Big(\big[(\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k})^{-1}\big]_{[3,3]}\Big)\Big)^{-1}. \tag{B.1}$$

The behavior of $D_s^{(12)}$ is compared for three types, namely, items with $\mathbf{a}_1 = (a, 0, 0)$, $\mathbf{a}_2 = (0, a, 0)$, and $\mathbf{a}_3 = (0, 0, a)$ for an arbitrary value of a. The first two types of items discriminate only with respect to one of the two intentional abilities while the third item discriminates

only with respect to the nuisance ability. (The earlier derivation of the criterion of D-optimality showed that items discriminating highly with respect to one ability are generally more informative than items discriminating with respect to multiple abilities. Hence, this choice.) Furthermore, it is assumed that the difficulty and guessing parameters of the three items are fixed and that $\hat{\boldsymbol{\theta}} = \mathbf{0}$. Therefore, $g(\boldsymbol{\theta}, \mathbf{a}, b, c)$ is fixed for all $\mathbf{a}$ and, without loss of generality, its value can be set equal to one.

The test information matrix for the choice of an item with $\mathbf{a}_1 = (a, 0, 0)$ as candidate item is

$$\mathbf{I}_{S_{k-1}} + \mathbf{I}_1 = \begin{bmatrix} C_{11} + a^2 & C_{12} & C_{13} \\ C_{12} & C_{22} & C_{23} \\ C_{13} & C_{23} & C_{33} \end{bmatrix},$$

where $C_{l_1 l_2}$ denotes element (l_1, l_2) of $\mathbf{I}_{S_{k-1}}$. Note that the test information matrices for the second and third types of items are similar to this matrix, except that a^2 should be added to the second and third instead of the first diagonal element.

For these three cases, $D_s^{(12)}$ can be written as

$$D_s^{(12)}(\mathbf{a}_1, b, c) = \frac{a^2 \det((\mathbf{I}_{S_{k-1}})_{[1,1]}) + \det(\mathbf{I}_{S_{k-1}})}{C_{33}},$$

$$D_s^{(12)}(\mathbf{a}_2, b, c) = \frac{a^2 \det((\mathbf{I}_{S_{k-1}})_{[2,2]}) + \det(\mathbf{I}_{S_{k-1}})}{C_{33}},$$

$$D_s^{(12)}(\mathbf{a}_3, b, c) = \frac{a^2 \det((\mathbf{I}_{S_{k-1}})_{[3,3]}) + \det(\mathbf{I}_{S_{k-1}})}{C_{33} + a^2},$$

respectively, where $\mathbf{I}_{S_{k-1}[l,l]}$ is the cofactor, that is, the submatrix obtained when omitting the lth row and lth column.

Observe that the criterion shows a linear increase with a^2 for the first two types of items while for the third type it increases asymptotically: $\det((\mathbf{I}_{S_{k-1}})_{[3,3]})$ as $a^2 \to \infty$. Hence, for a sufficiently large, $D_s^{(12)}$ always selects the item that discriminates with respect to one of the two intentional abilities. However, the third type of item is most informative, that is,

$$\max_{\mathbf{a}} D_s^{(12)}(\mathbf{a}, b, c) = D_s^{(12)}(\mathbf{a}_3, b, c)$$

when

$$\begin{aligned} &C_{33}\big(\det\big((\mathbf{I}_{S_{k-1}})_{[3,3]}\big) - \det\big((\mathbf{I}_{S_{k-1}})_{[1,1]}\big)\big) > \det(\mathbf{I}_{S_{k-1}}) \quad \text{with } 0 < a < a' \text{ and} \\ &C_{33}\big(\det\big((\mathbf{I}_{S_{k-1}})_{[3,3]}\big) - \det\big((\mathbf{I}_{S_{k-1}})_{[2,2]}\big)\big) > \det(\mathbf{I}_{S_{k-1}}) \quad \text{with } 0 < a < a'', \end{aligned} \tag{B.2}$$

where

$$a' = \sqrt{\frac{C_{33} \det((\mathbf{I}_{S_{k-1}})_{[3,3]}) - \det(\mathbf{I}_{S_{k-1}}) - C_{33} \det((\mathbf{I}_{S_{k-1}})_{[1,1]})}{\det((\mathbf{I}_{S_{k-1}})_{[1,1]})}},$$

$$a'' = \sqrt{\frac{C_{33} \det((\mathbf{I}_{S_{k-1}})_{[3,3]}) - \det(\mathbf{I}_{S_{k-1}}) - C_{33} \det((\mathbf{I}_{S_{k-1}})_{[2,2]})}{\det((\mathbf{I}_{S_{k-1}})_{[2,2]})}}.$$

From the conditions on $\mathbf{I}_{S_{k-1}}$ in (B.1), it can be concluded that the third type of item is most informative when the current information about the intentional abilities is large relatively to the current information about all abilities, i.e., $\det((\mathbf{I}_{S_{k-1}})_{[3,3]}) > \det(\mathbf{I}_{S_{k-1}})$. The result is explained

by the fact that although θ_3 is a nuisance ability, it does have an impact on the probability of answering the items correctly. Upon the selection of items that only test the intentional abilities, the sampling variance of the estimator of the nuisance ability will becomes large relative to those of the intentional abilities. At this point, an item that mainly tests the nuisance ability may produce a larger decrease of the variance of the intentional ability estimators than items that test these intentional abilities only.

A.1. θ_1 Intentional and θ_2 and θ_3 Nuisance Abilities

In this case, candidate item i_k is selected to maximize

$$D_s^{(1)}(\mathbf{a}_{i_k}, b_{i_k}, c_{i_k}) = \left(\left[(\mathbf{I}_{S_{k-1}} + \mathbf{I}_{i_k})^{-1}\right]_{(1,1)}\right)^{-1}, \tag{B.3}$$

which is the inverse of the sampling variance of the estimator of the intentional ability. We make the same assumptions as in the preceding case and are therefore able to set function g equal to one. For items with $\mathbf{a}_1 = (a, 0, 0)$, $\mathbf{a}_2 = (0, a, 0)$, and $\mathbf{a}_3 = (0, 0, a)$, $D_s^{(1)}$ can be written as

$$D_s^{(1)}(\mathbf{a}_1, b, c) = \frac{a^2 \det((\mathbf{I}_{S_{k-1}})_{[1,1]}) + \det(\mathbf{I}_{S_{k-1}})}{\det((\mathbf{I}_{S_{k-1}})_{[1,1]})},$$

$$D_s^{(1)}(\mathbf{a}_2, b, c) = \frac{a^2 \det((\mathbf{I}_{S_{k-1}})_{[2,2]}) + \det(\mathbf{I}_{S_{k-1}})}{\det((\mathbf{I}_{S_{k-1}})_{[1,1]}) + a^2 C_{33}},$$

$$D_s^{(1)}(\mathbf{a}_3, b, c) = \frac{a^2 \det((\mathbf{I}_{S_{k-1}})_{[3,3]}) + \det(\mathbf{I}_{S_{k-1}})}{\det((\mathbf{I}_{S_{k-1}})_{[1,1]}) + a^2 C_{22}},$$

respectively. The most informative item is

$$\max_{\mathbf{a}} D_s^{(1)}(\mathbf{a}, b, c)$$

$$= \begin{cases} D_s^{(1)}(\mathbf{a}_2, b, c) & \text{if } C_{33}^{-1} \det((\mathbf{I}_{S_{k-1}})_{[1,1]})(\det((\mathbf{I}_{S_{k-1}})_{[2,2]}) - \det((\mathbf{I}_{S_{k-1}})_{[1,1]})) > \det(\mathbf{I}_{S_{k-1}}) \\ & \text{with } 0 < a < a', \\ D_s^{(1)}(\mathbf{a}_3, b, c) & \text{if } C_{22}^{-1} \det((\mathbf{I}_{S_{k-1}})_{[1,1]})(\det((\mathbf{I}_{S_{k-1}})_{[3,3]}) - \det((\mathbf{I}_{S_{k-1}})_{[1,1]})) > \det(\mathbf{I}_{S_{k-1}}) \\ & \text{with } 0 < a < a'', \\ D_s^{(1)}(\mathbf{a}_1, b, c) & \text{otherwise} \end{cases}$$

where

$$a' = \sqrt{\frac{\det((\mathbf{I}_{S_{k-1}})_{[1,1]}) \det((\mathbf{I}_{S_{k-1}})_{[2,2]}) - (\det((\mathbf{I}_{S_{k-1}})_{[1,1]}))^2 - C_{33} \det(\mathbf{I}_{S_{k-1}})}{\det((\mathbf{I}_{S_{k-1}})_{[1,1]}) C_{33}}},$$

$$a'' = \sqrt{\frac{\det((\mathbf{I}_{S_{k-1}})_{[1,1]}) \det((\mathbf{I}_{S_{k-1}})_{[3,3]}) - (\det((\mathbf{I}_{S_{k-1}})_{[1,1]}))^2 - C_{22} \det(\mathbf{I}_{S_{k-1}})}{\det((\mathbf{I}_{S_{k-1}})_{[1,1]}) C_{22}}}.$$

The interpretation of this result is similar to the preceding case with two intentional abilities: when the sampling variance of the one of the estimator of one of the nuisance abilities becomes too large, it becomes beneficial to select an item that decreases it.

References

Berger, M.P.F., & Wong, W.K. (Eds.) (2005). *Applied optimal design*. London: Wiley.
Bernaerts, K., Servaes, R.D., Kooyman, S., Versyck, K.J., & Van Impe, J.F. (2002). Optimal temperature input designs for estimation of the square root model parameters: parameter accuracy and model validity restrictions. *International Journal of Food Microbiology*, *73*, 145–157.
Bloxom, B., & Vale, C.D. (1987). Multidimensional adaptive testing: An approximate procedure for updating. In *Meeting of the psychometric society*. Montreal, Canada, June.
Boughton, K.A., Yao, L., & Lewis, D.M. (2006). Reporting diagnostic subscale scores for tests composed of complex structure. In *Meeting of the national council on measurement in education*. San Francisco, CA, April.
Chang, H.-H. (2004). Understanding computerized adaptive testing: from Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *Handbook of quantitative methods for the social sciences* (pp. 117–133). Thousand Oaks: Sage.
Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.
Fan, M., & Hsu, Y. (1996). Multidimensional computer adaptive testing. In *Annual meeting of the American educational research association*. New York City, NY, April.
Lehmann, E.L. (1999). *Elements of large-sample theory*. New York: Springer.
Luecht, R.M. (1996). Multidimensional computer adaptive testing. *Applied Psychological Measurement*, *20*, 389–404.
McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monographs No. 15*.
McDonald, R.P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258–270). New York: Springer.
Owen, R.J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.
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.
Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412.
Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous items response data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.
Samejima, F. (1974). Normal ogive model for the continuous response level in the multidimensional latent space. *Psychometrika*, *39*, 111–121.
Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.
Segall, D.O. (2000). Principles of multidimensional adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–73). Boston: Kluwer Academic.
Silvey, S.D. (1980). *Optimal design*. London: Chapman & Hall.
Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the military testing association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.
Tanner, M.A. (1993). *Tools for statistical inference*. New York: Springer.
van der Linden, W.J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, *20*, 373–388.
van der Linden, W.J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412.
van der Linden, W.J. (2005). *Linear models for optimal test design*. New York: Springer.
van der Linden, W.J., & Glas, C.A.W. (Eds.) (2000). *Computerized adaptive testing: Theory and practice*. Boston: Kluwer Academic.
van der Linden, W.J., & Glas, C.A.W. (2007). Statistical aspects of adaptive testing. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 27. Psychometrics* (pp. 801–838). Amsterdam: North-Holland.
van der Linden, W.J., & Veldkamp, B.P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics*, *32*, 398–418.
Veldkamp, B.P., & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588.
Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A primer*. Hillsdale: Lawrence Erlbaum Associates.
Yao, L., & Boughton, K.A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*, 83–105.

Manuscript Received: 18 SEP 2007
Final Version Received: 6 NOV 2008
Published Online Date: 23 DEC 2008