

Making Fixed-Precision Between-Item Multidimensional Computerized Adaptive Tests Even Shorter by Reducing the Asymmetry Between Selection and Stopping Rules

Applied Psychological Measurement
2020, Vol. 44(7-8) 531–547
© The Author(s) 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0146621620932666

journals.sagepub.com/home/apm



Johan Braeken¹  and Muirne C. S. Paap^{2,3} 

Abstract

Fixed-precision between-item multidimensional computerized adaptive tests (MCATs) are becoming increasingly popular. The current generation of item-selection rules used in these types of MCATs typically optimize a single-valued objective criterion for multivariate precision (e.g., Fisher information volume). In contrast, when all dimensions are of interest, the stopping rule is typically defined in terms of a required fixed marginal precision per dimension. This asymmetry between multivariate precision for selection and marginal precision for stopping, which is not present in unidimensional computerized adaptive tests, has received little attention thus far. In this article, we will discuss this selection-stopping asymmetry and its consequences, and introduce and evaluate three alternative item-selection approaches. These alternatives are computationally inexpensive, easy to communicate and implement, and result in effective fixed-marginal-precision MCATs that are shorter in test length than with the current generation of item-selection approaches.

Keywords

computerized adaptive testing, fixed precision, item selection rules, variable length, multidimensional IRT

Tailoring a test to a specific respondent has been a popular approach for many decades. The idea of adaptive item selection can be traced back to the first intelligence tests, where the test would be terminated once the correct “mental age” could be determined with sufficient certainty. In the last decades, adaptive testing has become much more sophisticated, owing in part to advancements in information technology (IT) and the development of item response theory

¹University of Oslo, Norway

²University of Groningen, The Netherlands

³Oslo University Hospital, Norway

Corresponding Author:

Muirne C. S. Paap, The Nieuwenhuis Institute for Educational Research, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Rozenstraat 38, 9712 TJ Groningen, The Netherlands.

Email: m.c.s.paap@rug.nl

(IRT). Computerized adaptive tests (CATs)¹ are becoming increasingly popular, and the continuing development in IT goes hand-in-hand with the further refinement of IRT and CAT methodology.

An important line of research focuses on multidimensional CAT (MCAT), which allows for “borrowing” of information across dimensions. Adams et al. (2016) described two types of multidimensional item response models: within-item and between-item multidimensional models, which correspond to the “complex” and “simple” structures in factor analysis, respectively (W.-C. Wang & Chen, 2004). In the current study, we focus on between-item multidimensionality, where each item relates to one subdimension only; multidimensionality is expressed through the correlations among the latent dimensions (for a thorough primer on multidimensional IRT, see Reckase, 2009). These types of MCATs are a popular choice across various fields, both in simulation studies and operational MCATs (e.g., Frey & Seitz, 2011; Lee et al., 2019; Makransky & Glas, 2013; C. Wang et al., 2019). By acknowledging the multidimensional dependence structure, MCAT typically results in more efficient tests as compared with using separate unidimensional CATs (UCATs) per dimension—a finding which has been shown to hold under a wide range of conditions (Paap et al., 2019).

Although MCATs hold a lot of promise, they are associated with a number of additional challenges as compared with UCATs. One of these challenges concerns the relation between the item-selection rule and stopping rule—two crucial components in every CAT. If the stopping rule is based on fixed precision (rather than fixed length), it is typically defined in similar terms as the selection rule in a UCAT: both criteria are a function of measurement precision. This direct link is advantageous in case of fixed-precision CAT, since it results in optimally short efficient tests. With the advancement of MCAT, a number of selection rules have been developed (see, for example, Mulder & van der Linden, 2009). In many cases, these are seemingly straightforward adaptations of the rules used for UCATs. Yet, when one intends to stop in terms of a fixed precision per dimension, the selection rules currently used in the context of fixed-precision MCAT are not directly linked to the stopping rules: Selection is based on multidimensional precision, whereas stopping is based on marginal precision. As a result, the MCAT administration will be suboptimal and result in longer test lengths than necessary. This discrepancy has hitherto received little attention in the literature. In this article, we will discuss this selection-stopping asymmetry, and introduce three solutions.

The remainder of the article is structured as follows. First, it is briefly described how measurement precision is defined in the multidimensional case, and several popular selection rules are discussed. Second, the asymmetry between selection and stopping rules used for MCATs is described, and it is argued how this asymmetry can negatively affect test length. Third, alternative item-selection rules are introduced that are more closely linked to fixed marginal precision stopping in MCATs. Fourth, the different selection rules are illustrated and evaluated using two types of simulation studies: one based on an “ideal” bank, and one based on an empirical bank. Finally, the implications of our findings are discussed and recommendations are given.

Measurement Precision in the Multidimensional Case

For didactical reasons, consider the two-dimensional case ($Q = 2$). The test information function is defined as the Fisher information matrix that the observed response vector \mathbf{Y} contains about the to-be-estimated person parameters θ_1 and θ_2 under the item response model with likelihood function $L(\mathbf{Y}|\theta)$:

$$\mathcal{I}(\theta) = \begin{bmatrix} -E \left[\frac{\partial^2}{\partial \theta_1 \partial \theta_1} \log L(\mathbf{Y}|\theta) \right] & -E \left[\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log L(\mathbf{Y}|\theta) \right] \\ & -E \left[\frac{\partial^2}{\partial \theta_2 \partial \theta_2} \log L(\mathbf{Y}|\theta) \right] \end{bmatrix}.$$

Under regularity conditions, the inverse of the information matrix results in the covariance matrix of the person parameter estimates $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$:

$$\begin{aligned} \mathbf{VAR}(\hat{\theta}) &= \begin{bmatrix} \mathit{VAR}(\hat{\theta}_1) & \mathit{COV}(\hat{\theta}_1, \hat{\theta}_2) \\ & \mathit{VAR}(\hat{\theta}_2) \end{bmatrix} \\ &= \mathcal{I}(\theta)^{-1} \\ &= \frac{1}{|\mathcal{I}(\theta)|} \begin{bmatrix} -E \left[\frac{\partial^2}{\partial \theta_2 \partial \theta_2} \log L(\mathbf{Y}|\theta) \right] & E \left[\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log L(\mathbf{Y}|\theta) \right] \\ & -E \left[\frac{\partial^2}{\partial \theta_1 \partial \theta_1} \log L(\mathbf{Y}|\theta) \right] \end{bmatrix}. \end{aligned} \tag{1}$$

Equation 1 shows that the variance of one latent trait estimate depends on the information that is present on the latent trait estimate of the other dimension and the amount of interdependence in information on the latent traits as reflected by the determinant of the information matrix $|\mathcal{I}(\theta)|$; the latter plays the role of a common scaling factor for all elements in the covariance matrix.

In the *unidimensional* case, the information matrix reduces to a single cell and selecting the next item i_k from the available item pool \mathcal{I}_k to maximize information is equivalent to minimizing the standard error and increasing measurement precision:

$$\arg \max_{i_k \in \mathcal{I}_k} \mathcal{I}(\theta) \equiv \arg \min_{i_k \in \mathcal{I}_k} SE(\hat{\theta}^{(k)}).$$

Hence, for fixed-precision UCAT, the objective criterion of the selection rule is exactly equivalent to that of the stopping rule; this symmetry ensures an optimal adaptive test administration.

When extending this stopping rule to the *multidimensional* case, the standard error of *each of Q latent trait estimates*² is required to be less than or equal to a fixed threshold value δ :

$$\text{If } \forall q SE(\hat{\theta}_q) \leq \delta \Rightarrow \text{Stop(MCAT)}. \tag{2}$$

This is a straightforward extension of the stopping rule in fixed-precision UCAT, where only the precision of this one dimension is to be considered. The extension of the previously described UCAT item-selection rule to the multidimensional case is less straightforward, however, because items contribute information on more than one latent dimension at a time. For within-item multidimensionality this contribution is caused by items measuring multiple dimensions, whereas for between-item multidimensionality this is due to the dimensions being correlated. As a consequence of the multidimensionality, the objective criterion for item selection is formulated in terms of multivariate precision and inevitably an asymmetry between selection and stopping rule arises: Multivariate precision is used for selection, whereas marginal precision is used for stopping. This will be illustrated for two widely used objective criteria in the optimal design literature (Pukelsheim, 2006): the D- and A-optimality criteria.

Perhaps, the most popular choice of objective criterion for item selection in MCAT is the determinant of the information matrix, known as the D-optimality criterion:

$$\mathbf{D}\text{-optimality selection rule: } \arg \max_{i_k \in \mathcal{I}_k} |\mathcal{J}(\theta)| \quad (3)$$

In multivariate statistics, the determinant is also known as a generalized variance measure, and maximizing the determinant of the information matrix is equivalent to minimizing the generalized multivariate variance of the latent trait estimates:

$$\arg \max_{i_k \in \mathcal{I}_k} |\mathcal{J}(\theta)| \equiv \arg \min_{i_k \in \mathcal{I}_k} |\mathbf{VAR}(\hat{\theta})|.$$

This rule will select the candidate item that leads to the largest decrease in volume of the confidence ellipsoid of the latent trait estimates (Segall, 1996). This is quite close to the situation we had in UCAT, but the determinant is not directly proportional to the measurement precisions for the marginal dimensions as given by the $SE(\hat{\theta}_q)$ s in the MCAT stopping rule. Stopping in terms of the D-optimality criterion implies terminating the MCAT administration in terms of its multivariate precision, instead of the intended marginal precision for each dimension.

An alternative MCAT selection rule selects the candidate item that leads to the largest reduction in the sum of expected marginal standard error around the latent trait estimates. The objective criterion is specified in terms of the trace of the covariance matrix of the estimates; in the optimal design literature referred to as A-optimality:

$$\mathbf{A}\text{-optimality selection rule: } \arg \min_{i_k \in \mathcal{I}_k} \sum \text{diag}(\mathbf{VAR}(\hat{\theta})). \quad (4)$$

In multivariate statistics the trace is also known as a total variance measure. This selection rule is a direct function of the $SE(\hat{\theta}_q)$ s that are crucial for the MCAT stopping rule, but the summation to total variance gives it a compensatory nature. In contrast, the fixed-precision stopping rule used when all dimensions are of interest is non-compensatory: Each single dimension needs to be measured with a fixed level of precision.

Asymmetry Between Selection and Stopping Rule in Fixed-Precision MCAT

As mentioned earlier, in fixed-precision UCAT, there can be a one-to-one relation between the objective criterion used in the selection rule and in the stopping rule. This symmetry (see top half of Figure 1) ensures the optimality of the fixed-precision UCAT. In contrast, when the objective in a MCAT is to reach a fixed level of precision for each dimension, there is an asymmetry (see bottom half of Figure 1) between the objective criterion used in the selection rule and in the stopping rule. Decreasing the determinant or trace of the covariance matrix does not guarantee that each of the marginal standard errors would decrease (although the average standard error across dimensions is expected to decrease).

Yet, Mulder and van der Linden (2009) point out that item-selection rules using the D- or A-optimality criterion have a built-in minimax mechanism:

When the estimator of one of the abilities has a small sampling variance, they develop a preference for items highly informative about the other abilities. Hence, the next item will be most informative for dimensions that are lagging behind in measurement precision. 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. (p. 280)

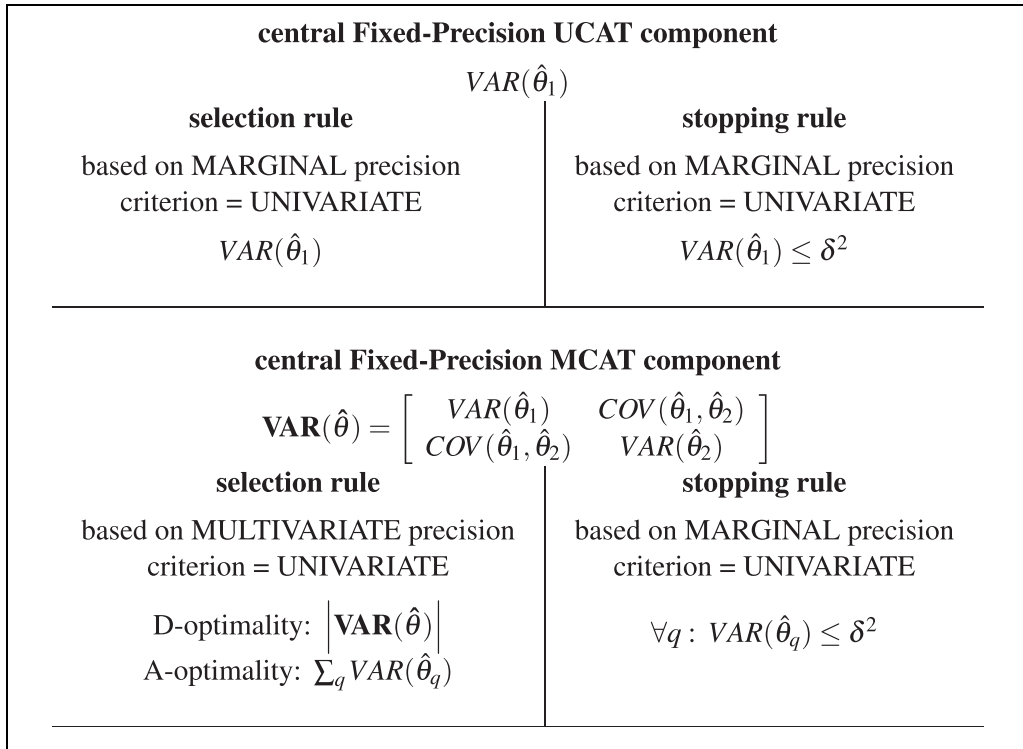


Figure 1. UCAT symmetry/MCAT asymmetry between selection and stopping rule.
 Note. UCAT = unidimensional computerized adaptive test; MCAT = multidimensional computerized adaptive test.

Their conclusion suggests that the asymmetry between selection and stopping rule might not have a big impact in practice. Yet, what is easily overlooked, and is implicitly touched upon in their discussion section, is that this minimax mechanism can only function properly when the CAT can draw items from a high-quality item bank, allowing for a wide choice among a large number of items for each dimension, with adequate targeting for the relevant trait range of the persons being tested. However, many empirical item banks suffer from information-range gaps and/or ceiling or floor effects; these issues may hamper the proper functioning of the minimax mechanism. Selecting items that are not well-matched to the examinee’s trait value could potentially result in an increase in measurement precision for a dimension whose measurement precision threshold δ was already met. In the worst-case scenario, this could lead to a continuous stream of selected items that are informative for increasing the *multivariate* precision, but not the *marginal* precision for the dimension whose marginal precision does not yet meet the δ threshold.

There is hardly any research that has focused on consequences of the asymmetry between the selection and stopping rule across wider trait ranges and less ideal item banks, although W.-C. Wang and Chen (2004) warned that current procedures might not necessarily stop at the potential minimum number of items and that further studies need to look into item-selection rules for fixed-precision MCATs.

We conjecture that the minimax tendency of the D-/A-optimality selection rules will function as expected for “ideal” item banks. However, for less-than-perfect item banks, we expect that the performance of the minimax tendency using current selection rules will be more erratic;

which may result in artificially long test lengths. We argue that there is a need for smarter selection rules that incorporate knowledge regarding which of the dimensions already meet the fixed marginal precision threshold.

In sum, a solution to the multivariate selection—marginal stopping asymmetry in fixed-precision MCAT where all dimensions are intentional—is needed that is both practical and intuitive. We propose a solution where the marginal precision is directly reflected in the definition of the objective criterion used for item selection. In the following, we will present three approaches to modify the widely used D- and A-optimality criteria for item selection.

Refining Item-Selection Rules for Fixed-Precision MCAT

Each of the three approaches outlined below is formulated in terms of a selection rule that applies to the posterior covariance matrix of the latent trait estimates and minimizes an objective function using a D- or A-optimality summary statistic.

Approach 1. Dynamically Restrict the Available Item Pool

The first approach we propose leaves the traditional item-selection criterion intact, but dynamically restricts the remaining item pool from which items can be selected. This can be done in two ways, which we will label “hard restriction” and “soft restriction.” Hard restriction implies that, once the precision threshold δ has been met for dimension q , the remaining items loading on that dimension can no longer be selected throughout the CAT, which virtually results in a new item pool \mathcal{S}_k^* (see, for example, Paap et al., 2018; Yao, 2013). Under soft restriction, only items can be selected that have a non-zero loading a_{id} on at least one dimension that does not yet meet the precision threshold δ at the given iteration during the CAT administration (see, for example, Paap et al., 2019). Hence, the item pool is defined as:

$$\mathcal{S}_k^* = \left\{ i_k \in \mathcal{S}_k : \exists_q \left[a_{i_k q} > 0 \ \& \ SE\left(\hat{\theta}_q^{(k-1)}\right) > \delta \right] \right\}. \quad (5)$$

Using this approach leaves the D- or A-optimality selection criterion intact:

$$\text{D-restricted selection rule: } \arg \min_{i_k \in \mathcal{S}_k^*} |\text{VAR}(\hat{\theta})|. \quad (6)$$

$$\text{A-restricted selection rule: } \arg \min_{i_k \in \mathcal{S}_k^*} \sum \text{diag}(\text{VAR}(\hat{\theta})). \quad (7)$$

Both the hard- and soft-restriction variants prohibit selection of items pertaining to dimensions for which the desired marginal measurement precision has already been reached, even if administering those items could have increased multivariate precision even further. Under soft restriction, any item that loads on a dimension for which precision has not yet been met is eligible for selection. Hard restriction would then in contrast only allow items that do not load on a dimension for which precision has already been met. Hence, the difference between the two variants is whether items measuring specific dimensions can become available again after an unintended increase of the standard error for the relevant dimension. Under hard restriction, the item pool restriction is permanent and not reversible, because it assumes an implied monotonic decrease of the marginal SEs during the CAT administration that is mathematically not strictly guaranteed. Note that both variants have been applied in specific studies as ad hoc solutions (see, for

example, Paap et al., 2018; Yao, 2013), but their performance has not been formally evaluated or compared with other approaches.

Approach 2. Dynamically Modify the Selection Criterion

In the second approach, dimensions for which the precision threshold δ is met are treated as nuisance dimensions (i.e., nuisance as in not relevant to be optimized). This approach can be seen as a dynamic alternative to the criterion proposed in Mulder and van der Linden (2009, p. 283). The nuisance status of a dimension is now a dynamic property throughout the CAT and a function of whether or not δ has been met for the relevant dimension:

$$\mathbf{VAR}(\hat{\theta})^{(W)} = \mathbf{W}^T \mathbf{VAR}(\hat{\theta}) \mathbf{W}, \tag{8}$$

where \mathbf{W} is a selection matrix of dimension $Q \times \left(\sum_q 1 \left[SE(\theta_q^{(k-1)}) > \delta \right] \right)$ that essentially filters out the dimensions for which δ has been reached. For instance, with $Q=3$ dimensions, at the start of the CAT, \mathbf{W} would be a three-dimensional diagonal matrix, yet as soon as δ is met for dimensions one and three, but not two, it would be reduced to $\mathbf{W}^T = [0, 1, 0]$, leading to $\mathbf{VAR}(\hat{\theta})^{(W)} = VAR(\hat{\theta}_2)$.

The filtered covariance matrix $\mathbf{VAR}(\hat{\theta})^{(W)}$ then forms the basis for a conventional item-selection rule:

$$\mathbf{D}\text{-filtered selection rule: } \arg \min_{i_k \in \mathcal{I}_k} \left| \mathbf{VAR}(\hat{\theta})^{(W)} \right|. \tag{9}$$

$$\mathbf{A}\text{-filtered selection rule: } \arg \min_{i_k \in \mathcal{I}_k} \sum \text{diag} \left(\mathbf{VAR}(\hat{\theta})^{(W)} \right). \tag{10}$$

The ‘‘nuisance’’ filtering is performed on the covariance matrix and not directly on the information matrix; using the information matrix would result in the information contributed by the ‘‘nuisance’’ dimensions being ignored, whereas marginal precision is a direct function of all dimensions (cf., Equation 1).

Approach 3. Focus the Selection Criterion Along Least Precisely Measured Dimension(s)

In the third approach, the covariance matrix is summarized in terms of a maximal direction.

For the D-optimality variant, the item that minimizes the largest eigenvalue λ_1 of the covariance matrix $\mathbf{VAR}(\hat{\theta})$ is selected:

$$\mathbf{D}\text{-max selection rule: } \arg \min_{i_k \in \mathcal{I}_k} \lambda_1 \left(\mathbf{VAR}(\hat{\theta}) \right). \tag{11}$$

This selection rule is virtually identical to the E-rule used in the optimal design literature, where the minimum eigenvalue of the information matrix is maximized. The traditional D-optimality criterion uses the determinant of the matrix—which is equivalent to the product of all eigenvalues. The largest eigenvalue and corresponding eigenvector represent the multivariate direction along which the least measurement precision is present. By specifically focusing on this direction, an item-selection procedure leading to shorter test lengths could be achieved. In the ideal case where all dimensions reach equal precision, all eigenvalues will be identical.

The A-optimality variant selects the item that minimizes the largest variance element in the covariance matrix:

Table 1. Overview of the 10 Different MCAT Item-Selection Rules in Terms of the Objective Criterion Being Minimized.

Optimality Approach/variant	arg min _{$i_k \in \mathcal{I}_k$} criterion	
	D Generalized variance	A Total variance
Vanilla Non-modified default	<i>D-vanilla</i> $ \mathbf{VAR}(\hat{\theta}) $	<i>A-vanilla</i> $\sum \text{diag}(\mathbf{VAR}(\hat{\theta}))$
Dynamic restriction Vanilla criterion with restricted item pool \mathcal{I}_k^* ; only items can be selected that have a non-zero loading a_{id} on at least one dimension that does not yet meet the precision threshold δ	<i>Hard-restricted D</i> <i>Soft-restricted D</i>	<i>Hard-restricted A</i> <i>Soft-restricted A</i>
Dynamic filter Dimensions for which the precision threshold δ is met are treated as nuisance dimensions	<i>D-filtered</i> $ \mathbf{VAR}(\hat{\theta})^{(\mathbf{W})} $	<i>A-filtered</i> $\sum \text{diag}(\mathbf{VAR}(\hat{\theta})^{(\mathbf{W})})$
Maximal direction Focus selection criterion along least precisely measured dimension(s)	<i>D-max</i> $\lambda_1(\mathbf{VAR}(\hat{\theta}))$	<i>A-max</i> $\max \text{diag}(\mathbf{VAR}(\hat{\theta}))$

Note. For hard-restricted, the item pool restriction is permanent; for soft-restricted, it is re-evaluated for each iteration and hence reversible. \mathbf{W} stands for a weight matrix coding for which dimension has not yet reached its fixed-precision threshold. λ_1 stands for the largest eigenvalue. MCAT = multidimensional computerized adaptive test.

$$\mathbf{A}\text{-max selection rule: } \arg \min_{i_k \in \mathcal{I}_k} \max \text{diag}(\mathbf{VAR}(\hat{\theta})). \quad (12)$$

Using a similar principle, C. Wang et al. (2012, Equation 11) formulated a fixed-precision stopping rule: stopping when the maximum standard error across dimensions is below a fixed-precision threshold.

Evaluating the Item-Selection Rules

In this section, the alternative selection criteria will be compared with their “vanilla” (i.e., classic) counterparts (i.e., Equations 3 and 4) using two simulation studies. An overview of the 10 MCAT selection rules, introduced in the preceding sections, is given in Table 1. Two item banks were used for the evaluation: an “ideal” bank and an empirical bank. MCATs with item selection based on the vanilla implementation of the D- and A-rule functioned as baseline conditions. As an additional check, a condition running Q separate UCATs was added.

CAT Algorithmic Settings

For both item banks, 10 MCATs were run that were similar in all algorithmic settings except for the item-selection rule. The CAT simulations were run in R (R Core Team, 2017) version 3.4 with the package mirtCAT (Chalmers, 2016) version 1.5.2. We customized the available item-selection rules in mirtCAT to match the selection rules described in this article (see R-code for customized item-selection rules available at <https://www.uv.uio.no/cemo/english/people/aca/>

johabrae/mirtcat-mcatitemselectionrules.r). For latent trait estimation, the maximum a posteriori (MAP) procedure was applied using a multivariate normal prior with 0-mean vector and correlation matrix \mathbf{R}_θ . The CATs were initialized by selecting the single most informative item for an average person in the population ($\theta = \mathbf{0}_{Q \times 1}$) and stopped once $SE(\theta_q < \delta) \forall q$, with $\delta = .387$. This value was chosen because it roughly corresponds to reliability values of .85; in clinical assessment, high reliability values are desirable, since the stakes are often high. See Raju et al. (2007) for more information on the topic of conditional reliability, including relevant equations.

In the UCAT reference condition, Q separate fixed-precision UCATs were run; the starting and stopping procedures were quasi-equivalent to those in the MCATs. Item selection was based on maximum Fisher information, and θ estimation was based on MAP with a univariate normal prior with 0-mean and unit-variance. Each UCAT was started and stopped independently from the other $Q - 1$ UCATs.

Evaluation: Performance Criteria

Feasibility of CAT administration was evaluated by examining whether the CATs under a given selection rule reached a *proper stop* (i.e., the required fixed precision δ on each dimension) for each individual simulated respondent (simulee). If the CAT did not reach a proper stop for a particular simulee, test length was set to the full-bank length to avoid artificial distortion of test length comparisons across selection rules (i.e., not reaching fixed precision implies item bank depletion).

Quality of CAT-based trait recovery was evaluated using the estimation bias per dimension. Bias was computed as the difference between the CAT-based $\hat{\theta}$ and the full-bank $\hat{\theta}$, with negative/positive values corresponding to under/overestimation. Bias diminishes as test length increases, and selection rules that result in shorter tests may therefore be at a natural disadvantage when it comes to bias. Therefore, the focus here was not on exact differences in bias, but rather on whether or not bias fell within an acceptable range for each rule under study.

Test length was evaluated using the total test length across the dimensions. Test length under each modified rule was contrasted to the test length under the D- and A-vanilla rule using a so-called landscaping technique (Navarro et al., 2004; Wagenmakers et al., 2004). Results were graphically inspected using scatterplots with each point representing a simulee's test length for a CAT administered using a specified item-selection rule compared with this simulee's test length for a CAT administration under the default vanilla selection rule; we will refer to these plots as landscape plots (see Supplemental Figure A1 for an illustration).

Evaluation: Item Sequence Characteristics

The selection rules were also evaluated in terms of characteristics of the resulting sequence of selected items. Similarity of item sequences was quantified for each simulee by computing the *relative overlap* in selected items $\left(\frac{\#(i(\text{rule}Y) \cap i(\text{rule}X))}{\#i(\text{rule}X)} \right)$ and the *relative divergence point* between the two sequences. The divergence point was defined as the first iteration in the CAT administration where the rule selected a different item than its comparison rule for simulee p . The relative divergence point equaled the divergence point divided by the resulting test length under the comparison rule. The number of times an item of a given dimension was followed by an item of the same dimension was used as an indicator for suboptimal performance of the mini-max mechanism. The item sequence characteristics will be treated as descriptors, not as performance quality indicators.

Study I: An Ideal Item Bank

A simulated item bank of $I = 260$ items (binary response categories), designed to be balanced and symmetric across a $Q = 4$ -dimensional latent space, was used as a basis for the first study. The bank was calibrated using a between-item multidimensional 2PL response model. For each latent dimension, a set of $J = I/Q = 65$ binary items was simulated as follows: The set of J item intercepts b_i were formed by combining an equal-step size sequence of length $\lceil 1.5J/5 \rceil$ in the interval $[2, 4]$ with an equal-step size sequence of length $J/5 - 1$ in the interval $[0, 2]$ and adding the mirror reflection of this combined series around a center value of 0. The item discriminations a_i were drawn from a uniform distribution in the interval $[0.8, 1.2]$. For each dimension, the simulated item set produced a relatively flat test information function, satisfying the required fixed-precision threshold $\delta = .387 \forall \theta \in [-2.5, 2.5]$ under a unidimensional 2PL model.

To cover a large area of the multidimensional latent space, a sample of $n = 10,000$ simulees was drawn, with the $\theta_{n \times Q}$ values drawn from a $Q = 4$ -dimensional continuous uniform grid with margins in the interval $[-2, 2]$. The prior correlation matrix \mathbf{R}_θ was set to have pairwise correlations equal to .7 for all dimensions. Given these settings, the obtainable marginal standard errors for the sample of simulees on each dimension based on the whole multidimensional item pool ranged between .324 and .343; these values are well below the required fixed-precision threshold $\delta = .387$.

Results

For most item-selection rule conditions, the CAT algorithm reached a proper stop for 100% of the $n = 10,000$ simulees, with exception of the two hard-restricted selection rules where for 70 simulees the CAT did not reach a proper stop. As seen in Figure 2, the A-vanilla rule performed slightly better than its D counterpart; this is in line with the criterion being mathematically closer to the marginal standard error than the D-vanilla rule. The vanilla rules' average test lengths corresponded to about 38% of the full-bank size, whereas combined test length for separate UCATs equaled 55% of the full-bank size. The D- and A-vanilla selection rules resulted in MCATs that were 6–8 items longer than the suggested alternative selection rules. Note that not only did the alternative rules result in shorter tests, the test length was also more homogeneous across simulees for the modified rules (cf. smaller interquartile range/box sizes). Although differences in test length between the vanilla rules and the UCAT condition were larger than the differences between the vanilla rules and their modified counterparts, the latter differences were still substantial. This is a surprising finding, given that an “ideal” itembank was used in this simulation study. With the bias per dimension being comparable between the A-/D-vanilla rules ($M = .00$, $SD = .20$) and the alternative versions (for each rule, $M = .00$, $SD = .20$), the shorter test length did not come at the expense of a substantial increase in estimation bias (for reference, UCAT: $M = .01$, $SD = .25$).

The landscape plots in Figure A2 (see Supplemental Appendix A) provide an overview of direct pairwise comparisons between the plain vanilla D-/A-rule and the suggested alternative selection rules. Test length was shorter for both vanilla rules as compared with the UCAT condition in 100% of the cases, with a minimum reduction of 14/16 items and a median reduction of 43/45 items, respectively. The A-vanilla rule resulted in shorter test length than the D-vanilla rule in 81% of the cases (median reduction of 2 items and maximum reduction of 14 items). The alternative selection rules resulted in shorter test length than the D-vanilla rule in 100% of the cases (median reduction of 8 items and maximum reduction of 24 items) and than the A-vanilla rule in 99% of the cases (median reduction of 6 items and maximum gain of 19 items), with the exception of the D-max rule for which the numbers equaled 97% and 91%, respectively. None

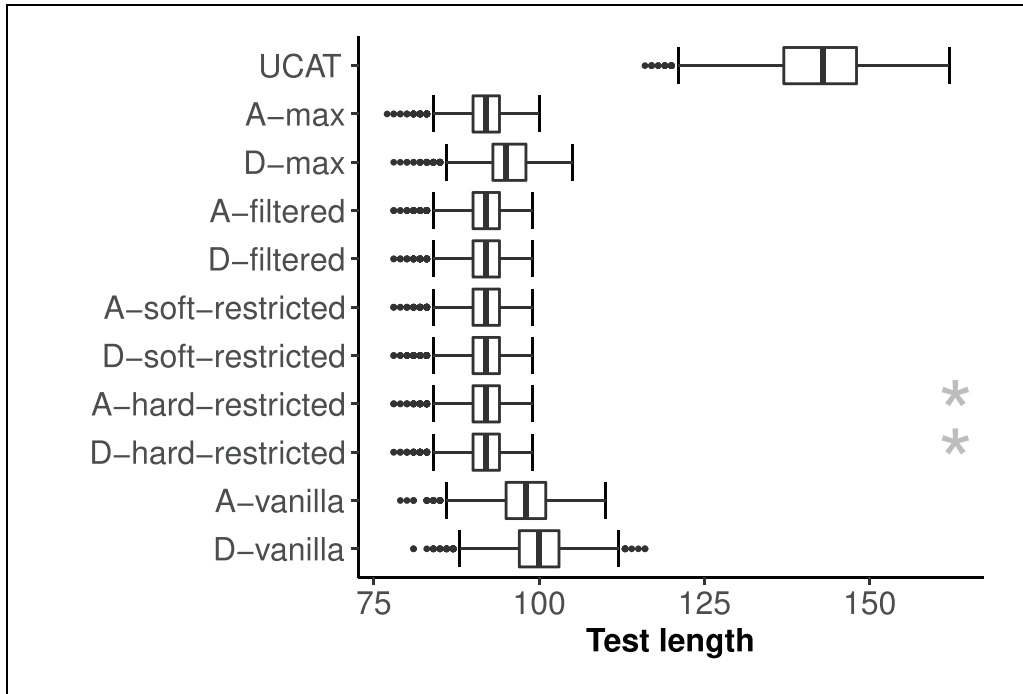


Figure 2. Ideal item bank: Distribution of CAT test length by item-selection rule.
 Note. CAT test length for $n = 10,000$ cases; the two gray asterisks represent the 70 and 71 improperly terminated cases for which the MCAT ran to bank depletion under the hard-restricted D- and A-rule, respectively. CAT = computerized adaptive test; MCAT = multidimensional computerized adaptive test.

of the alternative selection rules resulted in a longer test length than the D-/A-vanilla rules in any of the cases, with the exception for the D-max selection rule compared with the A-vanilla rule (1% of the cases).

The A-vanilla rule ended up selecting mostly the same items as the D-vanilla rule ($M = 98\%$; see Figure 3A). Out of the alternative selection rules, the D-max rule showed the highest degree of relative overlap in selected items with the D-/A-vanilla rules ($M = 95/97\%$). The relative overlap does not take into account the order of the items in the selection sequence during the MCAT. Hence, looking at the relative divergence point complements the picture (see Figure 3B). Although the D-vanilla rule showed high relative overlap in selected items with both D- and A-variants of the alternative selection rules, the relative divergence point confirmed the D-/A-family similarities, with the D-restricted and D-filtered variants' item sequences running in parallel (on average) with the D-vanilla rule for up to 83% of its test length and the A-restricted and A-filtered variants running in parallel (on average) with the A-vanilla rule for up to 87% of its test length. Divergence from the opposite family variants generally occurred at an early stage, after only 6%–7% of the test length (i.e., approximately 6 items). The median number of times that an item was followed by an item of the same dimension equaled 16 for the selection rules belonging to the D-family, compared with 9 times for the A-family. The max variants were the odd ones out, with a median of only 4 and 1 time for D-max and A-max.

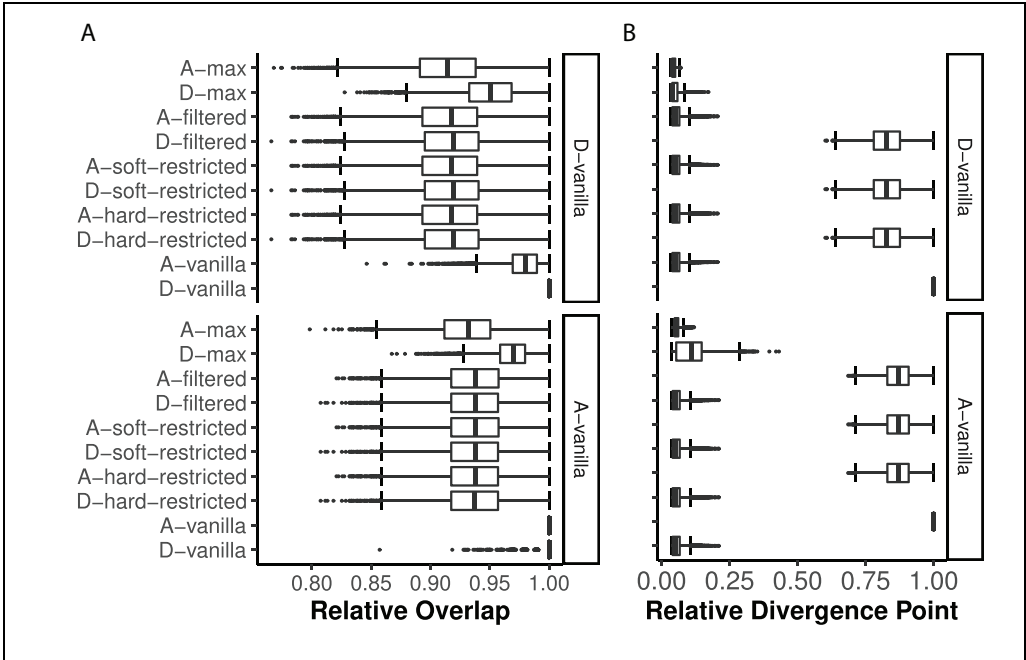


Figure 3. Ideal item bank: (A) Relative overlap and (B) relative divergence point of item sequences of alternative selection rules with the D-/A-vanilla rules.

Study II: A Real Item Bank Example

An empirical multidimensional item bank of $I = 194$ items (4 or 5 ordinal response categories), designed to measure different aspects of quality of life (Paap et al., 2018), was used as a basis for the second study. Four dimensions were measured: fatigue (50 items), disease-specific complaints (46 items), physical function (63 items), and social roles and activities (35 items). The bank was calibrated using a between-item multidimensional graded response model. Higher scores were indicative of higher quality of life for all dimensions. As is often seen in health measurement, all dimensions were highly positively correlated, and items had high discrimination parameters (see Supplemental Table B1). The threshold parameters covered a wide range for each dimension.

A sample of $n = 10,000$ simulees was drawn, with the $\theta_{n \times Q}$ values drawn from a $Q = 4$ -dimensional continuous uniform grid with margins in the interval $[-2, 2]$. The prior correlation matrix \mathbf{R}_θ was set to the estimated population correlation matrix. The marginal standard errors for each of the four dimensions were well below the fixed-precision threshold of $\delta = .387$ for all simulees if the full item bank would be administered ($SE(\hat{\theta}_q)$ range: $[0.09, 0.20]$).

Results

For all but two selection rules, the CAT algorithm reached a proper stop for 100% of the $n = 10,000$ simulees. For the D- and A-restricted selection rules where a hard restriction was imposed, this figure dropped to 81%–82%. The highly discriminating polytomous items with well-spread thresholds made it possible to achieve precise measurement with about two handfuls of well-selected items for each selection rule (see Figure 4). The average total test length was longest for the D-vanilla selection rule (12 items), followed by the combined UCATs (10 items),

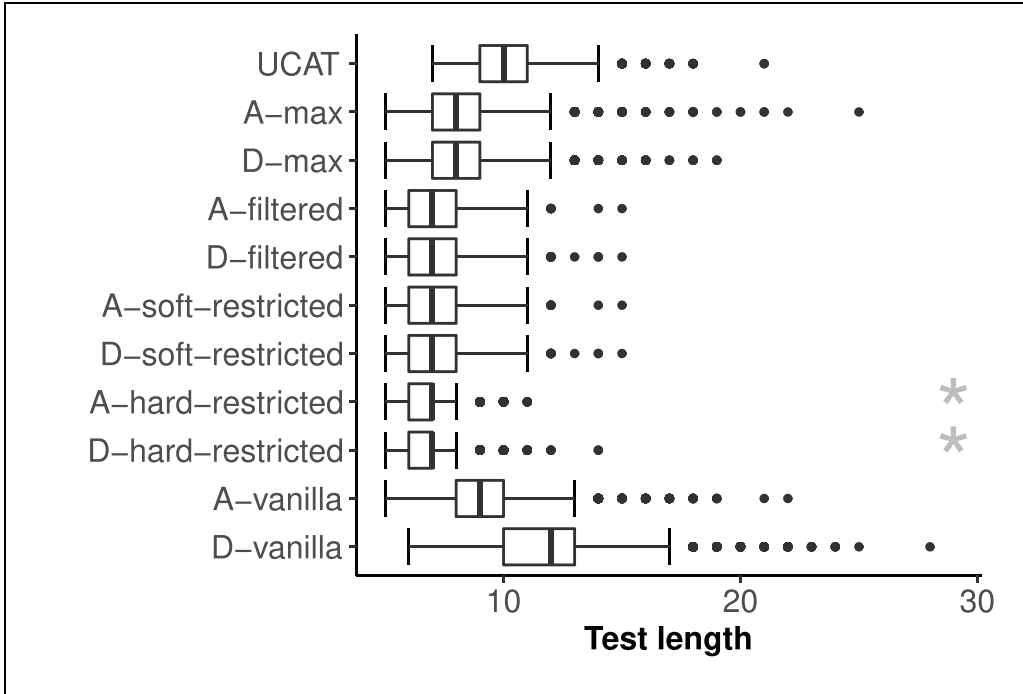


Figure 4. Empirical item bank: Distribution of CAT test length by item-selection rule.
 Note. CAT test length for $n = 10,000$ cases; the two gray asterisks represent the 1,789 and 1,863 improperly terminated cases for which the MCAT ran to bank depletion under the hard-restricted D- and A-rule, respectively. CAT = computerized adaptive test; MCAT = multidimensional computerized adaptive test.

A-vanilla rule (9 items), and both the max-variant selection rules (8 items). The shortest average test length was observed for the restricted and filtered selection rules (7 items). Although average bias remained negligible under all selection rules, the short test lengths coupled with the multivariate prior resulted in a larger variation in bias under the MCAT selection rules ($M = .03$, $SD = .51$) than under the UCAT condition ($M = .00$, $SD = .29$), with the exception of the D-vanilla rule ($M = .03$, $SD = .34$) which had a relatively longer test length (on average 3 extra items). As expected, this bias-precision trade-off applied mostly to simulees with relatively extreme θ -values, with an over-/underestimation bias for low/high θ -values due to prior-shrinkage.

The landscape plots in Supplemental Figure B1 provide an overview of direct pairwise comparisons between the plain vanilla D-/A-rule and the proposed alternative selection rules. The D-vanilla rule resulted in a shorter test length as compared with UCATs for 22% of the simulees, and in longer test length for 64% of the simulees. The latter observation is a realization of the worst-case consequence of the selection-stopping asymmetry, where a stream of selected items further reduces multivariate precision but not the marginal precision for the dimension that is not yet meeting the stopping criterion. For the A-vanilla rule, these figures were 62% and 17%, respectively. It being less affected by the asymmetry is consistent with the A-optimality selection criterion being somewhat closer to the stopping criterion than D-optimality. Furthermore, the alternative selection rules resulted in shorter test length than the D-vanilla rule for at least 97% of the simulees, and 81% for the A-vanilla rule, and never were longer than the combined UCATs test length. The exceptions were the D-max and A-max rules, where the

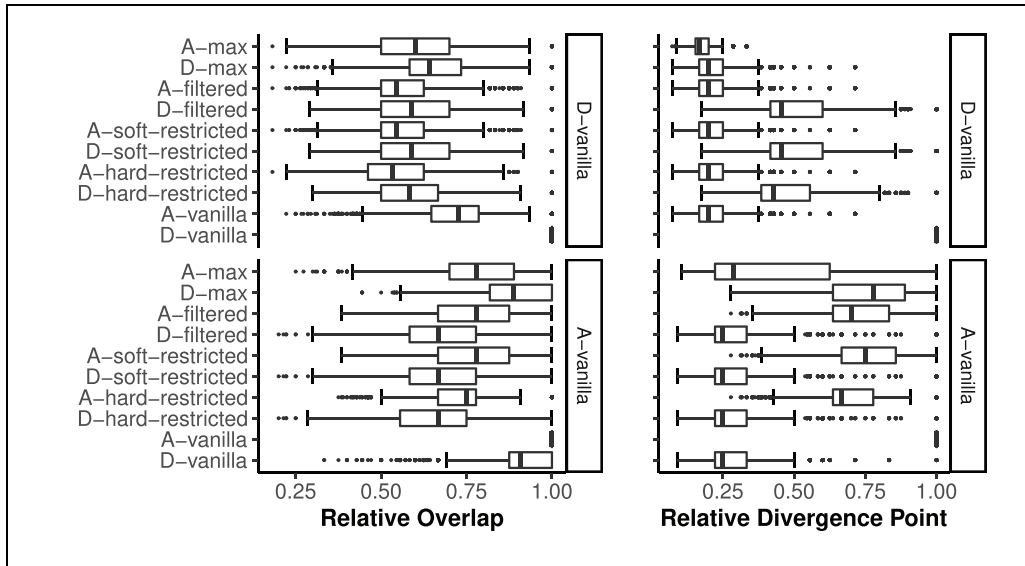


Figure 5. Empirical item bank: Relative overlap and divergence point of item sequences of alternative selection rules with the D-/A-vanilla rules.

numbers dropped to 92/91% and 60/65%, respectively, and that were also occasionally outperformed by the combined UCATs.

Although a moderate to high degree of relative overlap was found for all selection rules, regardless of A-/D-family, the relative divergence point results showed more differentiation (see Figure 5). The alternative rules were generally more similar (higher relative divergence point) to their respective vanilla rule than to the other rules. On average, the item sequences produced by the D-restricted and D-filtered rules were identical to those under the D-vanilla rule for up to 49% of its test length; this number increased to 70% for their A-variant counterparts. Divergence from opposite family variants generally occurred at an early stage: after only 20%–25% of the test length (i.e., approximately 2–3 items). The D-max rule was again an exception, diverging from its D-family at an early stage but from the A-family of selection rules only at a late stage. All selection rules had a median number of 1 time that an item was followed by an item of the same dimension, except for the vanilla and max variants where this was a median number of 0 times.

Discussion

In the case that all dimensions are intentional, we showed that test length for between-item fixed-precision MCATs can be decreased through modification of the traditional D- and A-optimality criteria for MCAT item selection by incorporating knowledge on which of the dimensions already meet the required fixed marginal precision stopping threshold. Three approaches addressing the asymmetry between the inherently multivariate nature of the selection criterion and the marginal nature of the fixed-precision stopping criterion were introduced: (a) dynamically restricting the available item pool \mathcal{S}_k (hard- and soft-restricted variants), (b) dynamically weighting the covariance matrix $\mathbf{VAR}(\hat{\theta})$ used as a basis of the selection criterion (filtered variants), and (c) focusing on the least precisely measured dimension (max variants).

We expected that under ideal circumstances (having a well-balanced, informative, and symmetric item bank), differences between the vanilla rules and the alternative rules would be small. Yet, the reduction in test length associated with using the modified selection rules was substantial for both the empirical bank and the ideal bank. Strikingly, the vanilla rules did not consistently outperform separate UCATs in terms of test length for the empirical bank. In fact, test length under the D-vanilla rule was longer than in the UCAT condition for a majority of the simulees. This finding underpins the importance of the need for an improved alignment of the item-selection rule with the stopping rule in fixed marginal precision between-item MCAT. Our findings imply that implementing one of the alternative selection rules rather than the vanilla selection rules can be expected to have a substantial impact on operational fixed marginal precision MCATs.

Not all proposed selection rules performed equally well. Although overall the hard-restricted selection rules resulted in shorter test length, CAT administration continued to bank depletion for a number of simulees. The hard-restricted rules virtually delete items pertaining to a specific dimension once the measurement precision on that dimension falls below the desired fixed-precision threshold. The latter is never re-evaluated throughout the remainder of the CAT administration: the *SE* is simply assumed to be monotonically non-increasing. However, this assumption is not mathematically supported; and especially at the start of a CAT, some variability in precision is expected across iterations in the multivariate setup (precision on one dimension also depends on the available information on the other dimensions, see Equation 1). Hence, when using a hard-restricted selection rule, there is a risk of items being removed from the active bank that might be needed later on in the CAT, which in turn may have substantial consequences for CAT feasibility. The number of CAT administrations for which this undesirable behavior occurred was rather low, but given that there are better options available, we would not recommend to pursue the hard-restricted selection variants any further nor implement these in operational CATs.

The D-max selection rule resulted in shorter CATs than the vanilla selection rules, but it did not perform as well as its competitors. Mulder and van der Linden (2009) also reported issues with this criterion (there listed as the E-rule) when used in fixed-length MCATs. They reported numerical instability and pointed out that the objective criterion did not map 1-to-1 on sampling variance for all dimensions (note that the latter feature was exactly why it was a reasonable candidate in our context). A similar conclusion as for the hard-restricted variants applies; given that there are better options available, it may not be advisable to implement the D-max selection rule in operational CATs.

Overall, the filtered and soft-restricted selection rules performed very favorably. Our findings suggest that these rules could be considered as the preferred choice for operational fixed-precision MCATs when all dimensions are intentional. Both filtered and soft-restricted rules are computationally inexpensive, easy to communicate, and easy to implement, with the filtered variants requiring a dynamically updated weighting matrix and the soft-restricted variants requiring a dynamically updated available item pool. Note that in the case of between-item multidimensionality, both approaches result in equivalent item-selection behavior.

If one would have to choose between the two vanilla rules, we suggest the A-vanilla rule should be favored since it consistently outperformed its D-rule counterpart in terms of test length; furthermore, it shows a higher degree of mathematical similarity with the fixed-precision stopping criterion.

The empirical example bank used in this study consisted of highly informative and discriminating polytomous items, which resulted in extremely short test lengths, especially for the

MCAT selection rules. These “ultra-short” test lengths had an adverse side-effect: The variation in bias was somewhat larger for the MCAT selection rules due to the influence of the multivariate prior, especially affecting simulees with latent trait combinations that have low probability of occurring given the prior. In such instances, the prior will pull the θ -estimates of the different dimensions closer together and to the center, which may not be a desirable effect for these types of θ patterns. If one wants to make sure that bias is within an acceptable range for all persons, it may be advisable to couple fixed-precision selection with a minimum number of items, as suggested by Babcock and Weiss (2012).

We recognize that item selection is just one aspect of the CAT machinery. As illustrated by the differences between the two item bank scenarios used in our study, it is important to keep in mind that efficiency and other qualities of measurement depend to a very large extent on the specific goals and item bank properties of the end user. This being said, our results have clearly illustrated that aligning the selection rules used in MCAT with the intended measurement purpose (measuring each dimension with a specific level of precision) can have a considerable impact in terms of performance by ameliorating the adverse effects of the asymmetry between the multivariate nature of the selection criterion and the marginal nature of the fixed-precision stopping criterion. Using the filtered or soft-restriction selection rules, which incorporate knowledge on which of the dimensions already meet the required fixed-precision threshold, can be expected to result in shorter test lengths for fixed marginal precision MCATs.


Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a FRIPRO Young Research Talent grant for the second author (Grant No. NFR 286893), awarded by the Research Council of Norway.

ORCID iDs

Johan Braeken  <https://orcid.org/0000-0002-2119-3222>

Muirne C. S. Paap  <https://orcid.org/0000-0002-1173-7070>

Supplemental Material

Supplementary material is available for this article online.

Notes

1. For historical background and theoretical foundations underlying CAT, see, for example, Chang (2015).
2. In principle, the required level of precision δ could be set separately per dimension, for instance, depending on their respective importance in the particular application (van der Linden, 1996) or on what is operationally feasible given the item bank.

References

- Adams, R. J., Wilson, M., & Wang, W.-C. (2016). The multidimensional random coefficients multinomial logit model. *Journal of Statistical Software*, *71*(5), 1–39.
- Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*, 1–18.
- Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–39.
- Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20.
- Frey, A., & Seitz, N.-N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the Programme for International Student Assessment. *Educational and Psychological Measurement*, *71*(3), 503–522.
- Lee, Y., Lin, K., & Chien, T. (2019). Application of a multidimensional computerized adaptive test for a clinical dementia rating scale through computer-aided techniques. *Annals of General Psychiatry*, *18*, Article 5.
- Makransky, G., & Glas, C. A. W. (2013). The applicability of multidimensional computerized adaptive testing for cognitive ability measurement in organizational assessment. *International Journal of Testing*, *13*(2), 123–139.
- Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*(2), 273–296.
- Navarro, D., Pitt, M., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*(1), 47–84.
- Paap, M. C. S., Born, S., & Braeken, J. (2019). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement*, *43*(1), 68–83.
- Paap, M. C. S., Kroeze, K. A., Glas, C. A. W., Terwee, C. B., van der Palen, J., & Veldkamp, B. P. (2018). Measuring patient-reported outcomes adaptively: Multidimensionality matters! *Applied Psychological Measurement*, *42*(5), 327–342.
- Pukelsheim, F. (2006). *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics.
- Raju, N. S., Price, L. R., Oshima, T., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*(3), 169–180.
- R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software manual]. <https://www.R-project.org/>
- Reckase, M. (2009). *Multidimensional item response theory*. Springer-Verlag.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354.
- van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, *4*(20), 373–388.
- Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*(1), 28–50.
- Wang, C., Chang, H.-H., & Boughton, K. A. (2012). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*(2), 99–122.
- Wang, C., Weiss, D. J., & Shang, Z. (2019). Variable-length stopping rules for multidimensional computerized adaptive testing. *Psychometrika*, *84*, 749–771.
- Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*(5), 295–316.
- Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*(1), 3–23.