Psychological Test and Assessment Modeling, Volume 56, 2014 (4), 348-367

Constrained Multidimensional Adaptive Testing without intermixing items from different dimensions

Ulf Kroehne[1], *Frank Goldhammer*[2] *& Ivailo Partchev*[3]

Abstract

Multidimensional adaptive testing (MAT) can improve the efficiency of measuring traits that are known to be highly correlated. Content balancing techniques can ensure that tests fulfill requirements with respect to content areas, such as the number of items from various dimensions (target rates). However, content balancing does not restrict the order in which items are selected from dimensions. If multiple dimensions are measured with MAT, intermixing items from different dimensions might invalidate properties of those items, which are known from calibration studies without mixed item content. To avoid this, the known correlations between traits can be used to increase efficiency of the ability estimation only without intermixing items from different dimensions. In this simulation study, MAT allowing items to be intermixed between dimensions is compared to Constrained MAT (CMAT) that does not allow intermixing items between dimensions for items with between-item multidimensionality. As expected, MAT achieved the greatest reliability for equal target rates; however, CMAT with items administered in a pre-specified order dimension by dimension was not disadvantageous for unequal target rates.

Keywords: multidimensional adaptive testing, content balancing, item response theory, multidimensional scoring, intermixing items

[1] *Correspondence concerning this article should be addressed to:* Ulf Kroehne, PhD, German Institute for International Educational Research, Schloßstraße 29, 60486 Frankfurt am Main, Germany; email: kroehne@dipf.de

[2] German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany; Centre for International Student Assessment (ZIB), Frankfurt/Main, Germany

[3] Cito, Arnhem, Netherlands

Introduction

Several tests measuring different but highly correlated abilities often are administered together as a test battery in one session. For instance, in the Psychological Service of the German Federal Employment Agency unidimensional computerized adaptive tests (CATs) are administered separately to measure intelligence and are based on various item types in order to obtain measures reflecting different facets and a broad representation of the intended latent construct. If multiple correlated dimensions such as skills, abilities or competencies are measured, multidimensional ability estimation (e.g., Bloxom & Vale, 1987; Brandt & Duckor, 2013) and multidimensional adaptive testing (MAT, Luecht, 1996; Segall, 1996) improve measurement efficiency (Frey & Seitz, 2009; Wang, Chen & Cheng, 2004)[4]. Overall, two of the potential advantages of MAT over multiple unidimensional CATs can be distinguished (Li & Schafer, 2005): a) the incorporation of correlations among abilities into the ability estimation and b) a more efficient item selection from the combined multidimensional item pool. In this paper practical considerations of MAT are taken into account and the relative importance of both benefits in the development of a MAT from existing CATs is highlighted.

Item selection and content balancing in MAT

Procedures for item selection in MAT are developed to optimize a psychometric criterion which is either the accuracy of a test taker's multidimensional ability estimate or classification decision (see, e.g., van der Linden & Pashley, 2010). Item selection optimizing a particular psychometric criterion does not necessarily restrict item selection with respect to content-related criteria between or within dimensions. Hence, if specific attributes of the items must be accounted for to render test assemblies among test takers comparable, advanced procedures for constrained item selection are required. If content requirements are necessary (for instance if each test is required to be composed by comparable proportions of particular item types), such methods extend the psychometric criterion for item selection and are known as content balancing techniques (see, e.g., van der Linden, 2005). Content balancing techniques are employed to ensure that individual tests satisfy additional constraints, for example, a specific number of items for each of several content areas (target rates).

General approaches developed for CAT such as shadow testing can be applied as a content balancing technique for MAT as well (e.g., Li & Schafer, 2005; Veldkamp & van der Linden, 2002) and combined with additional constraints such as those related to item exposure (Diao, van der Linden & Yen, 2011). Moreover, specific approaches to content balancing such as the Modified Constrained CAT method (MCCAT; Leung, Chang & Hau, 2003) and the Maximum Priority Index (MPI; Cheng & Chang, 2009, Frey, Cheng & Seitz, 2011) were developed for MAT.

[4] Note, potential weaknesses of ability estimation in multidimensional IRT models need to be considered for individual-level diagnostics (see, e.g., Hooker, Finkelman & Schwarzman, 2009; van Rijn & Rijmen, 2012).

Intermixing items from different dimensions in MAT

In MAT with between-item multidimensionality (e.g., Reckase, 2009) each item can be clearly assigned to one dimension (also known as multidimensional item response theory models with a simple structure). However, the order in which items from different dimensions are selected by the MAT algorithm is determined solely by the item selection criterion (e.g., the determinant of the posterior information matrix), resulting in intermixed administrations of items from various dimensions.

Content balancing techniques can be used in MAT not only to control content-related criteria but also to restrict the number of items administered from each dimension (Frey, Seitz & Kroehne, 2013). This can render individual tests being comparable in terms of the number of items administered from each dimension. Tests become comparable among test takers when target rates used for content balancing reflect the assigned dimension of an item as content area, comparable tests allow reporting scores for (sub-)dimensions. Using content balancing techniques to control the number of items from each dimension also allows comparing composite scores of test takers if computed as a function of the dimension-specific ability estimates. That is, content balancing techniques are necessary to maintain comparability for a changeover from multiple CATs to MAT when the mixture of components is important for a composite score, computed as the (weighted) sum of the dimension-specific estimates. However, neither MAT with nor without procedures developed for content balancing (such as shadow testing, MCCAT and MPI) constrain the order in which items from different content areas are selected. Without additional constraints, items in MAT are selected from different dimensions. However, switching back and forth between dimensions must be considered when MAT is discussed as a further development of CATs for existing item pools. Little is known about the potential effects of items that are calibrated without being intermixed with those from various dimensions. From a psychological point of view, intermixing items from different dimensions might affect the test results in a multitude of ways. Potential sources of the various effects include additional cognitive demands of responding to items from different dimensions in a varying order, adhering to potential motivational and meta-cognitive effects, as well as item context effects, typically not studied in item calibration designs (see, e.g., Yousfi & Böhme, 2012). Hence, we conducted a simulation study to determine the need for further investigation into the potential effects of intermixing items from different dimensions. If performance on MAT is substantially lower when intermixing items is avoided by the suggested Constrained MAT (CMAT) algorithm, it would be necessary to study effects of empirically intermixing items among dimensions prior to conducting MAT for existing test batteries developed as multiple CATs. However, if we can show that CMAT is comparable to MAT with respect to measurement efficiency under reasonable conditions, MAT without intermixing items from different dimensions might be considered a useful and even preferable testing procedure.

Aims

Measurement efficiency of MAT is increased by the following components: a) incorporation of the latent correlation into the provisional and final ability estimation and b) more efficient item selection from a combined item pool. Avoiding intermixing items from different dimensions by the test algorithm inhibits the latter. The aim of this paper is to disentangle the effect of a) and b) for an existing test battery of CATs that also could be administered as MAT or CMAT. To do so, three research questions have been formulated. The first two questions deal with the increase in efficiency compared to the baseline condition of separate CATs, that is, compared to independent unidimensional adaptive tests with three independent unidimensional provisional and final ability estimations.

1) How much efficiency can be gained by using a final multidimensional ability estimation of the responses obtained from the administration of separate, independent, unidimensional CATs?
2) How much efficiency can be gained by administering unconstrained MAT with multidimensional provisional ability estimations, followed by separate final unidimensional ability estimations of the responses obtained from the unconstrained MAT?

The regular MAT conditions, both with MPI or MCCAT and multidimensional ability estimation, are expected to result in the most efficient test design. The third research question deals with the decrease in measurement efficiency of a newly developed CMAT procedure when compared to MAT with different content balancing techniques.

3) How much efficiency is lost by CMAT, i.e., from constraining MAT to a prespecified non-mixed order of dimensions?

To answer this question, we investigated whether intermixing items from different dimensions can be avoided while still profiting from the combined administration as MAT with content balancing. Van der Linden (2010) claimed that constraining item selection in MAT to only one dimension at a time (i.e., CMAT) would be disadvantageous from a psychometric point of view. However, the relative amount of efficiency lost due to this restriction is unknown.

It is important to note that MAT-MPI, MAT-MCCMAT and CMAT are identical with respect to the provisional and normally final multidimensional ability estimation. A comparison between unidimensional and multidimensional final ability estimation for fixed form tests within a hierarchical Bayesian approach was found by de la Torre and Patz (2005) for simulated items of a mathematics test. We extend their results by investigating the degree to which efficiency can be improved with final multidimensional ability estimation compared to multiple CATs by considering a modified MAT procedure without intermixing items from different dimensions, and by incorporating item parameters from an operational used item pool. In addition, we study MAT-MPI, MAT-MCCAT and CMAT with unidimensional final ability estimation.

Method

We conducted a simulation study in which we compared the different test algorithms using a real item pool of a given test battery and simulated responses. The number of items administered from each dimension was the same across all conditions. For the baseline conditions with separate CATs this was achieved by using test length as a termination criterion. For MAT the number of items from each dimension was specified as a target rate for the content balancing techniques MPI and MCCAT.

Frey, Cheng and Seitz (2011) developed the MPI as an extension of the Priority Index (Cheng & Chang, 2004) for MAT. The MPI can be employed with various item selection procedures (Yao, 2013) and is implemented by pre-multiplying a priority index with the criterion used for item selection, for example, the volume or the determinant of the information (Segall, 1996). The priority index is computed after each item selection as a weight from the intended target rates of the content constraint and the percentage of already selected items at a particular step of the test. Items are simultaneously selected from all dimensions in MAT using MPI, and content balancing is controlled by weighting items according to their assigned content area.

Instead of weighting simultaneously all the items from the different content areas, an alternative two-step approach presented by Leung, Chang and Hau (2000) as MCCAT can be used as content balancing for MAT. MCCAT was developed to eliminate the predictability of the sequence in which items from different content areas are selected (Leung, Chang & Hau, 2003). First, a cumulative probability distribution for each of the particular content areas is constructed by comparing the realized number of items and the desired target rate for each content area. One content area is sampled according to this cumulative probability distribution. Second, item selection is restricted to the subset of items from the sampled content area, that is, to the sampled content area, which is not sufficiently represented in a test taker's individual test.

Constrained MAT (CMAT) without intermixing items from different dimensions

When assigning items to one of the dimensions as the content area, MPI, MCCAT and possible precursors developed for content balancing (e.g., Constraint CAT by Kingsbury & Zara, 1989) do result in MAT with intermixing items from different dimensions. However, a content balancing technique can be implemented resulting in administration without intermixing items from different content areas. For this, item selection of a MAT is restricted to one content area at a time until the target rate of the content area is completely achieved. Then, the current content area is changed to the next content area and items are selected from the new content area. For this approach the order of content areas is fixed in advance, meaning that items from different content areas are selected in a fixed, pre-specified order. The resulting CMAT procedure is very similar to MAT with MCCAT. However, the order of content areas is defined in advance rather than being

randomly sampled proportionally to the cumulative probability after each administered item.

For the first dimension CMAT equals unidimensional CAT, but for the subsequent dimensions the estimation of the provisional ability benefits from multidimensional IRT, that is, previous responses to preceding dimensions will result in more precise ability estimates for the current dimension. Accordingly, the provisional ability for the second and following dimensions will converge faster to the true ability, and the increased precision of the ability estimation will result in a selection of items that are likely more informative at the true ability.

The observed measurement efficiency of the CMAT procedure might be order-dependent, as the results depend on the properties of the item pool (Reckase, 2010), which might differ between dimensions, as well as on the test length for each dimension (the target rates). In contrast to sequences of CATs optimized for individual test takers (van der Linden, 2010), the sequence of dimensions for CMAT will be known in advance, which allows test takers to be informed about upcoming tasks, for instance, in the instruction for the test battery. Substantive reasons and psychometric criteria can be considered when choosing a specific order of a CMAT test battery. In our simulation study, we consider all possible sequences to find the most efficient order of dimensions averaged across all test takers.

Test battery and item pool

The conditions of the simulation study were chosen to ensure comparability of the adaptive testing practices of the German Armed Forces and the German Federal Employment Agency. Item parameters from the three unidimensional adaptive tests measuring *inductive reasoning* with Raven matrices-like items (M), *numerical reasoning* with arithmetic items (N), and *verbal analogies* items (V) derived from the calibration sample (Storm, 1999; Hornke, 1999; Hornke, 2000; Hornke, Etzel & Küppers, 2000) were combined as a three dimensional 2-PL model (Birnbaum, 1968) for the simulation study. For each item i the probability of a correct response ($U_{ij}=1$) for simulee $j=1,\ldots,N$ is assumed to be a logistic function of three latent abilities $\boldsymbol{\theta}_j=\left(\theta_{j1},\theta_{j2},\theta_{j3}\right)$ and a set of item parameters $\mathbf{a}_i$ and b_i:

$$P\left(U_{ij}=1|\boldsymbol{\theta}_j,\mathbf{a}_i,b_i\right)=\frac{\exp\left(\mathbf{a}_i^{'}\left(\boldsymbol{\theta}_j-b_i\mathbf{1}\right)\right)}{1+\exp\left(\mathbf{a}_i^{'}\left(\boldsymbol{\theta}_j-b_i\mathbf{1}\right)\right)}.$$

The loading of item i on the different dimensions is represented by the 1×3 discrimination vector $\mathbf{a}_i^{'}$, which is assumed to have only one non-zero element for each item, i.e., a simple structure without cross loadings (between-item multidimensionality). The difficulty of item i is given by b_i. A summary of the item pool is presented in Table 1.

Table 1:
Item Pool of the Test Battery Used for the Simulation Study

Dimension	Number of Items	Discrimination			Difficulty		
		Mean	Median	SD	Mean	Median	SD
Matrices (M)	455	1.461	1.358	0.653	1.326	1.125	1.772
Numerical reasoning (N)	287	1.491	1.436	0.642	-0.272	-0.372	1.729
Verbal analogies (V)	254	1.402	1.324	0.535	-1.054	-0.987	1.497

The empirical correlations between the three dimensions as found in the calibration sample were disattenuated through dividing them by the root of the product of the assumed to be constant reliabilities. These disattenuated correlations were also used to define the multivariate prior distribution for the MAT as well as a true multivariate ability distribution for the data generation (.71 for the correlation between M and N, .68 for the correlation between M and V, and .69 for the correlation between N and V). A composite score, defined as the sum of the theta estimates for each dimension, was used to combine the dimension-specific results.

Procedure

Simulated theta values were sampled from a multivariate normal distribution with correlations Φ corresponding to the disattenuated correlations and unit variances (as well as a mean vector $\boldsymbol{\mu}$ of zero). Responses were generated according to the true thetas assuming perfect fit of the 2-PL model for all items in the item pool. Using these responses, we simulated the following testing methods, both with unidimensional and multidimensional final theta estimation:

(i) Multiple independent CATs,
(ii) MAT with two content balancing approaches (MPI and MCCAT), and
(iii) CMAT as described above.

The first research question (RQ1) can be answered from the comparison between unidimensional and multidimensional final theta estimations of the multiple CATs, the testing method simulated as condition (i). Comparing testing methods (i) and (ii) with unidimensional final theta estimates will help us answer the second research question (RQ2). Finally, the third research question (RQ3) will be answered by comparing testing methods (ii) and (iii), both with unidimensional and multidimensional final theta estimates.

Two sets of target rates were examined for all test algorithms, resulting in a total of 30 items each (equal target rates, 10 M items, 10 N items and 10 V items vs. unequal target rates, 8 M items, 10 N items and 12 V items). The number of items for each dimension under the unequal target rates condition was chosen to compensate for the different quality of the item pool in the three dimensions.

Table 2:
Summary of the Simulated Testing Methods with Unidimensional and Multidimensional Final Theta Estimation

Final Theta Estimation		Testing Method	Order of Dimensions	Number of Sequences	Provisional Theta Estimation	Research Question
Unidimensional Estimation	i)	Multiple CATs	Irrelevant	D*(D-1)	Unidimensional Estimation	RQ2, RQ1
	ii)	MAT-MPI / MAT-MCCAT	Intermixed item presentation	1	Multidimensional Estimation	RQ3, RQ1
	iii)	CMAT	Must be specified in advance	D*(D-1)	—	RQ3, RQ1
Multidimensional Estimation	i)	Multiple CATs	Irrelevant	D*(D-1)	Unidimensional Estimation	RQ1
	ii)	MAT-MPI / MAT-MCCAT	Intermixed item presentation	1	Multidimensional Estimation	RQ3
	iii)	CMAT	Must be specified in advance	D*(D-1)	—	RQ3

Bayes modal estimation with a standard normal prior was used to estimate the ability for CAT, and items with the highest Fisher information were selected without additional constraints. The multidimensional Bayes modal estimation was used for MAT based on a multivariate normal prior distribution with mean vector $\boldsymbol{\mu}$ and correlations Φ as used for data generation. The Bayesian approach of item selection according to the largest decrement in the volume of the confidence ellipsoid as described by Segall (1996) was implemented (D-optimality).

Data generation and the different test algorithms used in the simulation study were implemented in proprietary software. Each individual test administration was simulated using a theta starting value of zero (CATs) or a vector of zeros (MAT and CMAT). In total, $R = 500$ replications each with a sample size of $N = 1000$ test takers were simulated.

Evaluation criteria

Before measurement efficiency of the test algorithms are reported on, we compare the *bias* of the theta estimates in each dimension as well as the bias of the resulting composite score over all three dimensions. The average of the difference between the true theta and the estimated theta for each dimension, as well as for the true composite score and the estimated composite score is computed for each replication with the following formula:

$$Bias = \frac{\sum_{k=1}^{N}\left(\hat{\theta}_k - \theta_k\right)}{N},$$

where $\hat{\theta}_k$ is the estimated theta or the estimated composite score for simulee k and θ_k is the known theta value or value of the composite score for simulee k. The mean bias, that is the average of the biases over the 500 replications for each condition, will be reported. The *mean squared error* (MSE) is evaluated to assess the accuracy of theta estimates for each dimension as well as for the composite score:

$$MSE = \frac{1}{N}\sum_{k=1}^{N}\left(\hat{\theta}_k - \theta_k\right)^2.$$

The average over all replications as described for the mean bias will be reported for each condition. Comparing the MSE of the baseline condition and the different test algorithms which incorporate multidimensionality allows considering *relative efficiency*, defined in line with de la Torre and Patz (2005) as the MSE of separate CATs over the MSE of the alternative methods:

$$RE = \frac{MSE_{\text{Separate CATs}}}{MSE_{\text{Alternative Method}}}.$$

A ratio greater than 1.0 indicates that the algorithm incorporating the multidimensionality into ability estimation and/or item selection is more efficient compared to the results obtained from three independent CATs. Finally, to determine the practical relevance of the achieved increase in measurement efficiency, we will report on the squared correlation between the true thetas and the corresponding estimates from the simulation, as well as the squared correlation between the true composite score and the estimated composite score as a measure of the *reliability* of the different methods (Allen & Yen, 1979).

Results

Bias

A small negative bias was observed for the composite score under all conditions (on average -0.019 over all conditions with equal target rates and -0.017 for all conditions with unequal target rates). No systematic differences between the various methods were found and the bias of the composite score as well as the bias for each dimension with multidimensional estimation was slightly smaller than that of the unidimensional estimation (see Table 3).

Mean Squared Error

A MSE of 0.302 and 0.304 was achieved for the composite score by administering separate unidimensional CATs with either equal or unequal target rates (see Table 4). Using multidimensional ability estimation of the responses to items selected as separate CATs with provisional unidimensional theta estimation, the MSE was reduced to 0.267 and 0.269 for the composite score and a similar reduction was observed for the ability estimates of each of the dimensions (RQ1). The MSE of the composite score for simulated algorithms using MAT combined with the unidimensional final ability estimation also decreased for equal and unequal target rates (RQ2, from 0.285 to 0.267 for equal target rates and 0.269 for unequal target rates). No differences between MAT-MPI (0.256 / 0.257) and MAT-MCCAT (0.257) in terms of MSE were observed. The MSE for the CMAT procedure was found to be comparable to MAT-MPI and MAT-MCCAT for the composite score for unidimensional and multidimensional final theta estimation (RQ3). Table 4 also provides MSE for the three dimensions, which shows that for unequal target rates, MSEs are more comparable between dimensions.

Relative efficiency

A comparison of the efficiency of the test algorithms to the administration of all tests of the battery as multiple CATs is presented in Figure 1. For the composite score, the relative efficiency of multidimensional ability estimation was 1.130 for equal and 1.127 for unequal target rates. This is larger than the relative efficiency of using combined MAT

Table 3:
Bias of the Composite Score and Ability Estimates for Different Test Algorithms

Final Theta Estimation Method	Testing Method	Sequence of Dimensions	Equal Target Rates				Unequal Target Rates			
			Composite Score	M (10)	N (10)	V (10)	Composite Score	M (8)	N (10)	V (12)
Unidimensional Estimation	Multiple CATs		-0.0209	-0.0049	0.0090	0.0168	-0.0168	-0.0053	0.0080	0.0141
	MAT-MPI		-0.0201	-0.0038	0.0095	0.0144	-0.0178	-0.0041	0.0084	0.0135
	MAT-MCCAT		-0.0202	-0.0042	0.0094	0.0150	-0.0176	-0.0049	0.0085	0.0141
	CMAT	M-N-V	-0.0182	-0.0049	0.0085	0.0146	-0.0159	-0.0057	0.0078	0.0138
		M-V-N	-0.0186	-0.0049	0.0088	0.0147	-0.0158	-0.0057	0.0080	0.0135
		N-M-V	-0.0202	-0.0040	0.0096	0.0146	-0.0176	-0.0047	0.0086	0.0138
		N-V-M	-0.0211	-0.0040	0.0096	0.0155	-0.0177	-0.0048	0.0086	0.0140
		V-M-N	-0.0214	-0.0043	0.0088	0.0169	-0.0172	-0.0051	0.0081	0.0142
		V-N-M	-0.0225	-0.0041	0.0097	0.0169	-0.0179	-0.0049	0.0086	0.0142
Multidimensional Estimation	Multiple CATs		-0.0184	-0.0028	0.0076	0.0136	-0.0158	-0.0029	0.0068	0.0119
	MAT-MPI		-0.0182	-0.0019	0.0080	0.0120	-0.0171	-0.0018	0.0072	0.0117
	MAT-MCCAT		-0.0180	-0.0023	0.0081	0.0122	-0.0168	-0.0024	0.0073	0.0119
	CMAT	M-N-V	-0.0167	-0.0029	0.0075	0.0121	-0.0158	-0.0032	0.0070	0.0119
		M-V-N	-0.0171	-0.0029	0.0079	0.0121	-0.0156	-0.0032	0.0073	0.0115
		N-M-V	-0.0179	-0.0021	0.0079	0.0121	-0.0166	-0.0022	0.0070	0.0119
		N-V-M	-0.0187	-0.002	0.0079	0.0128	-0.0167	-0.0023	0.0070	0.0120
		V-M-N	-0.0194	-0.0022	0.0081	0.0135	-0.0165	-0.0025	0.0073	0.0117
		V-N-M	-0.0204	-0.0020	0.0088	0.0136	-0.0172	-0.0023	0.0077	0.0118

Note. M = Matrices, N = Numerical reasoning, V = Verbal analogies.

Table 4:
Mean Squared Error of the Composite Score and Separate Ability Estimates for Different Test Algorithms

Final Theta Estimation Method	Testing Method	Sequence of Dimensions	Equal Target Rates				Unequal Target Rates			
			Composite Score	M (10)	N (10)	V (10)	Composite Score	M (8)	N (10)	V (12)
Unidimensional Estimation	Multiple CATs		0.302	0.066	0.083	0.105	0.304	0.081	0.083	0.09
	MAT-MPI		0.285	0.063	0.080	0.101	0.285	0.076	0.080	0.088
	MAT-MCCAT		0.285	0.063	0.079	0.102	0.285	0.075	0.079	0.088
	CMAT	M-N-V	0.286	0.066	0.079	0.101	0.290	0.081	0.079	0.087
		M-V-N	0.287	0.066	0.078	0.101	0.290	0.081	0.078	0.088
		N-M-V	0.286	0.062	0.083	0.101	0.286	0.074	0.083	0.087
		N-V-M	0.286	0.061	0.083	0.101	0.284	0.073	0.083	0.088
		V-M-N	0.286	0.062	0.078	0.105	0.284	0.075	0.078	0.090
		V-N-M	0.287	0.061	0.079	0.105	0.283	0.074	0.079	0.090
Multidimensional Estimation	Multiple CATs		0.267	0.061	0.074	0.093	0.269	0.073	0.074	0.081
	MAT-MPI		0.256	0.059	0.073	0.091	0.257	0.070	0.072	0.080
	MAT-MCCAT		0.257	0.059	0.072	0.091	0.257	0.069	0.072	0.080
	CMAT	M-N-V	0.257	0.061	0.072	0.090	0.260	0.073	0.072	0.080
		M-V-N	0.258	0.061	0.071	0.091	0.261	0.073	0.071	0.080
		N-M-V	0.258	0.058	0.074	0.091	0.258	0.069	0.074	0.080
		N-V-M	0.257	0.058	0.074	0.091	0.257	0.068	0.074	0.080
		V-M-N	0.257	0.058	0.071	0.093	0.257	0.069	0.071	0.081
		V-N-M	0.257	0.058	0.072	0.093	0.256	0.068	0.072	0.081

Note. M = Matrices, N = Numerical reasoning, V = Verbal analogies.

with unidimensional estimation (1.061 for equal and 1.064 unequal target rates). As expected, the highest relative efficiency for the composite score under the condition with equal target rates was observed for MAT-MPI with multidimensional ability estimation and equal target rates (1.179). Moreover, the relative efficiency of MAT-MPI was not higher than that of MAT-MCCAT for unequal target rates (1.180).

For the CMAT procedure and equal target rates, relative efficiency for the composite score was slightly lower for all possible sequences of the three dimensions than that of MAT-MPI, both for unidimensional and for multidimensional estimation. For multidimensional estimation the relative efficiency of one condition of CMAT was at least as high as for MAT-MCCAT (V-M-N). For unequal target rates, three possible sequences of the three dimensions had a higher relative efficiency than MAT-MPI or MAT-MCCAT (N-V-M, V-M-N and V-N-M).

For unequal target rates, the relative efficiency for the composite score achieved by a particular sequence of dimensions in CMAT was related to the increase in the MSE of a dimension at a particular position (see Table 4). For each dimension, the MSE under

Figure 1:
Relative Efficiency of the Composite Score by Different Test Algorithms

multidimensional theta estimation was smaller for a particular dimension if this dimension was administered later in the CMAT sequence. However, this increase was small for the MSE of dimension V (MSE equals 0.080 for position 2 or 3 vs. 0.081 for position 1), larger for dimension N (MSE equals 0.071 for position 3, 0.072 for position 2 and 0.074 for position 1), and largest for dimension M (MSE equals 0.073 for position 1, 0.069 for position 2 and 0.068 for position 1). In line with these benefits of the MSE for particular dimensions, we found the greatest relative efficiency for the composite score for the sequence V-N-M (unidimensional estimation 1.072 / multidimensional estimation 1.187), followed by V-M-N (1.07 / 1.182) and N-V-M (1.068 / 1.182), which were more efficient than N-M-V (1.063 / 1.175), M-N-V (1.047 / 1.165) and M-V-N (1.046 / 1.165).

Reliability

As expected for aggregated measures, only slight differences were observed for the resulting reliability of the different algorithms (Wang & Chen, 2004). The reliabilities, obtained from the squared correlation of the true and estimated values, differed significantly between unidimensional and multidimensional estimation for each test algorithm

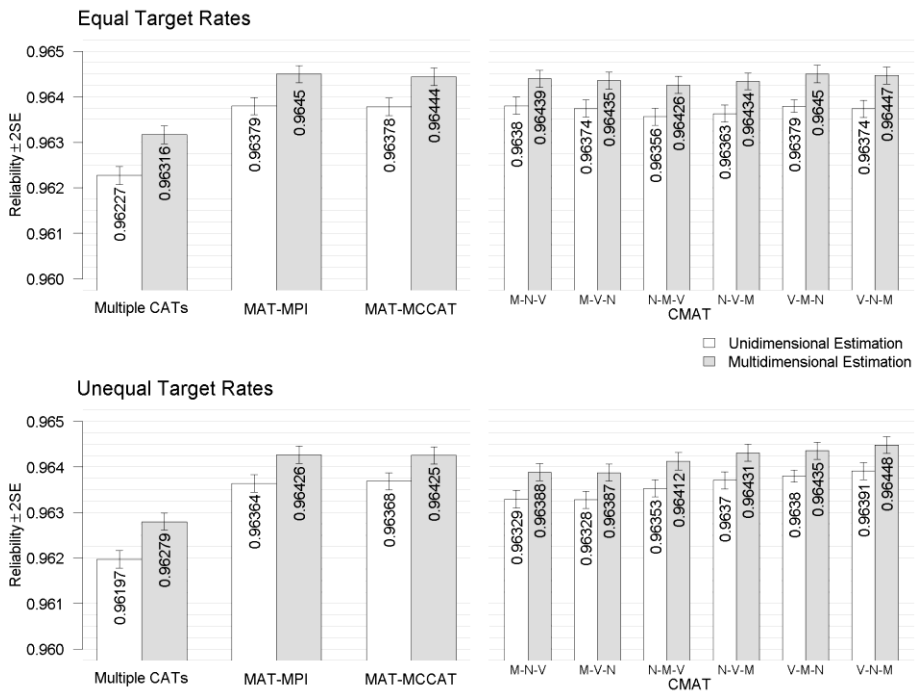

Figure 2:
Differences in Reliability of the Composite Score by Different Test Algorithms

(see upper part in Figure 2 for equal target rates). The reliability of unidimensional estimation of responses gathered with MAT (MPI, MCCAT or CMAT) was significantly greater than the reliability of the composite score of responses gathered with multiple CATs for unidimensional estimation (equal/unequal target rates: 0.96227/0.96197) and multidimensional estimation (equal/unequal target rates: 0.96316/0.96279). Reliability did not differ significantly between MAT-MPI and MAT-MCCAT, as well as CMAT, regardless of multidimensional ability estimation, or unidimensional ability estimation. The descriptive order of the composite scores' overall reliability was similar to the relative efficiencies of the test algorithms, with MAT-MPI yielding the greatest reliability. However, the reliability observed for CMAT for the sequences V-M-N and V-N-M were comparably high.

A similar finding was observed for unequal target rates (see lower part of Figure 2). Similar to the condition with equal target rates, reliability for some of the possible sequences of the CMAT procedure was significantly less than that for MAT. However, for other sequences (e.g., V-N-M) reliability was even greater than with MAT with one of the studied content balancing algorithms.

Discussion

The relative efficiency gained by using multidimensional theta estimation of the responses obtained from the administration of separate CATs was 1.130 (and 1.127 for unequal target rates). These values are comparable to the relative efficiencies of 1.15 and 1.24 reported by de la Torre & Patz (2005) for two and five dimensions using a multidimensional expected a posteriori method (MEAP) for ability estimation of fixed form tests (30 items, correlation of 0.70). With respect to RQ 1, we found that measurement efficiency could be improved as well by incorporating known correlations between traits into multidimensional theta estimation for CATs, but the increase was slightly smaller than the reported improvement in measurement efficiency for fixed form tests (at least for equal target rates).

The relative efficiency of using a MAT-MPI procedure in test administration followed by separate unidimensional theta estimation (RQ2) was found to be less (equal/unequal target rates: 1.061/1.064) than the increase in efficiency that was observed for multivariate theta estimation of multiple CATs (equal/unequal target rates: 1.13/1.127). For equal target rates, the relative efficiency for unidimensional estimation (see white bars in the upper part of Figure 1) of unconstrained MAT-MPI or MAT-MCCAT (1.061) were greater than that for CMAT (ranged between 1.054 and 1.59). This also led us to conclude that incorporating the correlation between traits for multidimensional ability estimation has a greater impact on measurement efficiency than optimizing item selection as MAT (followed by unidimensional theta estimation). However, it should be noted that the reliability for multivariate theta estimation of multiple CATs was less than the reliability based on the unidimensional estimation from items administered as MAT. Although this difference might not be astonishing from a statistical point of view, it has some practical relevance for tests with between-item multidimensionality. Whenever

multidimensional ability estimation is in doubt, for instance, because solving items from one dimension might potentially (even negatively) affect the ability estimate of another dimension (see Footnote 4), a possible practical strategy that can be suggested based on our research is to use CMAT with multidimensional item selection and multidimensional ability estimation during test administration but to report unidimensional final ability estimates. Doing so will avoid to incorporate information from other dimensions into the reported scores but will increase the reliability more than using multidimensional ability estimates based on multiple unidimensional CATs.

Using a constrained MAT procedure which avoids intermixing items from different dimensions was less optimal than unconstrained MAT for equal target rates (RQ3). However, for the specific item pool and in particular for unequal target rates, pre-selected sequences of the dimension can achieve a relative efficiency almost as good as that of unconstrained MAT with respect to the composite score. If the best sequence of dimensions is determined in a pre-operational simulation, CMAT is comparable to MAT-MCCAT for the composite score in terms of MSE and relative efficiency as well as in terms of reliability. The theoretically derived advantage of MAT-MPI over MAT-MCCAT is not significant for the real item pool used in the simulation study under all conditions, most obviously for the conditions with unequal target rates. In other words, we found that a higher relative efficiency with respect to the composite score could be achieved by using the CMAT approach, compared to MAT-MPI and MAT-MCCAT. As mentioned, this is because the relative efficiency of CMAT depends on the domain-specific properties of the item pool. However, in our simulation study the unequal target rates were chosen to compensate for the properties of the item pool, which is a limitation for possible interpretations of the findings from our simulation study. We can report that the efficiency of the different CMAT sequences for the composite score was related to the increase in the MSE of a dimension at a particular position. In addition, we can conclude that for item selection according to the studied D-optimality, the CMAT procedure implemented with at least one particular sequence of dimensions achieves an equal or better MSE than MAT-MPI and MAT-MCCAT for the composite score as well as for all dimension-specific scores (see Table 3). A more systematic investigation into the effect of different properties of the item pool and the amount of latent correlation between the dimensions on the relative performance of different sequences of CMAT is beyond the scope of this paper and subject to further research.

Our results are relevant for testing programs whenever multiple correlated traits such as sub-dimensions are assessed and combined in a single score as a (weighed) composite score, that is, when a test battery of multiple tests is operationally administered to achieve a particular content coverage. Multidimensional adaptive testing or alternative approaches to incorporating the known correlation of traits could be used to increase measurement efficiency.

We compared three different adaptive multidimensional procedures to a baseline condition with multiple CATs and found that the CMAT procedure, that is MAT without intermixing items from different dimensions, could be more efficient than multidimensional ability estimation of separately administered CATs only. This means that CMAT might be worth considering as a testing procedure whenever intermixing items from

different dimensions should be avoided for multidimensional adaptive testing. We also found that CMAT resulted in lower MSEs and higher reliabilities than multiple separate CATs did if a final unidimensional theta estimation was used. This finding is of particular importance if properties of a multidimensional theta estimation are questioned with respect to an individual level interpretation of this ability estimation.

To apply our results of increasing measurement efficiency to an existing test battery, it is necessary to find the best pre-specified order of dimensions for the CMAT and this requires a pre-operational simulation study, probably supplemented with additional criteria such as item exposure rates. Alternatively, applying the principles of adaptive testing at the test level (instead of the level of items or testlets) as suggested by van der Linden (2010) would be necessary in order to use the most efficient sequence of tests. Note, however, that in van der Linden's approach the order of dimensions would differ among test takers according to the given responses, and test takers could not be instructed or briefed in advance about the up-coming (sub-)dimensions of the test, as is possible with separate (adaptive) tests and the CMAT approach studied in this paper. In future research involving optimized test batteries the effect of varying sequences of domains among test takers might be considered, which occur, for instance, when booklet designs are used in large-scale assessments.

A limitation of this simulation study is that we analyzed the composite score as a derived quantity of dimension-specific theta estimates; however, we did not consider the composite score for item selection. Specific item selection algorithms for composite scores exist for MAT (van der Linden, 1999; Yao, 2012) and were recently combined with the MPI (Yao, 2013). Future research might investigate MAT without intermixing of dimensions for the estimation of composite scores, might address specific properties of the item pool, different levels of the correlation between dimensions, and might also include further evaluation criteria, such as item pool utilization and the conditional bias of theta estimation.

The focus of this research project was psychometric properties of a test battery administered as MAT or CMAT. We compared the relative efficiency of different test algorithms incorporating multidimensionality, and we conclude that under specific conditions the efficiency of the CMAT procedure is comparable to that of MAT with either MPI or MCCAT. Thus, using CMAT avoids unpredictable negative effects due to intermixing items among dimensions as in unconstrained MAT, whereas comparable levels of reliability can be reached. Beyond this psychometric result, using CMAT with content balancing restricted to a pre-determined order of dimensions also might yield further side effects as opposed to multiple unidimensional CATs, for instance, with respect to test takers' motivation and to position effects, which should be studied in subsequent empirical research.

Acknowledgements

This research was supported partly by grant KR 3994/1-3 from the German Research Foundation (DFG) in the Priority Programme ''Models of Competencies for the Assess-

ment of Individual Learning Outcomes and the Evaluation of Educational Processes'' (SPP 1293).

We are grateful to Lutz Hornke for his help in making the data available for this study.

References

Allen, M. & Yen, W. (1979). Introduction to Measurement Theory. CA: Brooks/Cole Publishing Company.

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

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

Brandt, S., & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling*, 55, 148-161.

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369-383.

de la Torre, J. & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30, 295-311.

Diao, Q., van der Linden, W. J. & Yen, S. J. (2011). Exposure control using item-ineligibility probabilities in multidimensional computerized adaptive testing with shadow test. Retrieved October 3, 2012, from http://www.ctb.com/img/pdfs/raExposureControlCompAdaptive.pdf

Frey, A., Cheng, Y., & Seitz, N. N. (2011, April). Content Balancing with the Maximum Priority Index Method in Multidimensional Adaptive Testing. Paper presented at the 2011 meeting of the National Council on Measurement in Education (NCME), New Orleans, LA, USA.

Frey, A. & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, 35, 89-94.

Frey, A., Seitz, N. N., & Kroehne, U. (2013). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Eds.), Research on PISA. Dodrecht: Springer.

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, 74, 419-442.

Hornke, L. F. (1999). Benefits from computerized adaptive testing as seen in simulation studies. *European Journal of Applied Psychology*, 15, 91-98.

Hornke, L. F. (2000). Item Response Times in Computerized Adaptive Testing. *Psicológica*, 21, 175–189.

Hornke, L.F., Etzel, S. & Küppers, A. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. [Construction and evaluation of an adaptive figural matrices test]. *Diagnostica*, 46, 182–188.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359-375.

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2000, April). *Content balancing in stratified computerized adaptive testing designs.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Leung, C. K., Chang, H., & Hau, K. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, 2 (5). Available from http://www.jtla.org

Li, Y. H. & Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement*, 42, 245-269.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389-404.

Reckase, M. D. (2009). Multidimensional Item Response Theory. Dordrecht: Springer.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, 52, 127-141.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.

Storm, E. G. (1999). Computeried adaptive testing in the Bundeswehr. Retrieved September 4, 2012, from http://iacat.org/sites/default/files/biblio/st99-01.pdf

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). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, 42, 283-302.

van der Linden, W. J. (2010). Sequencing an adaptive test battery. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 103-119). New York: Springer.

van der Linden, W. J. & Pashley, P. J. W. (2010). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 103-119). New York: Springer.

van Rijn, P. & Rijmen, F. (2012). *A Note on Explaining Away and Paradoxical Results in Multidimensional Item Response Theory* (Research Report No. RR-12-13). Princeton, NJ: Educational Testing Service.

Veldkamp, B. P. & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 64, 575-588.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 28, 450-480.

Wang, W. C., Chen, P.H., & Cheng, Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116-136.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and Applications, *Psychometrika*, 77, 495-523.

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules, *Applied Psychological Measurement*, 37, 3-23.

Yousfi, S., & Böhme, H. (2012). Principles and procedures of considering context effects in the development of calibrated item pools: Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling, 54*, 366-396.