Article

Measuring Patient-Reported Outcomes Adaptively: Multidimensionality Matters!

Applied Psychological Measurement
1–16

DOI: 10.1177/0146621617733954
journals.sagepub.com/home/apm


Muirne C. S. Paap[1,2], Karel A. Kroeze[3], Cees A. W. Glas[3], Caroline B. Terwee[4], Job van der Palen[3,5], and Bernard P. Veldkamp[3]

Abstract

As there is currently a marked increase in the use of both unidimensional (UCAT) and multidimensional computerized adaptive testing (MCAT) in psychological and health measurement, the main aim of the present study is to assess the incremental value of using MCAT rather than separate UCATs for each dimension. Simulations are based on empirical data that could be considered typical for health measurement: a large number of dimensions (4), strong correlations among dimensions (.77-.87), and polytomously scored response data. Both variable- ($SE < .316$, $SE < .387$) and fixed-length conditions (total test length of 12, 20, or 32 items) are studied. The item parameters and variance–covariance matrix $\mathbf{\Phi}$ are estimated with the multidimensional graded response model (GRM). Outcome variables include computerized adaptive test (CAT) length, root mean square error (RMSE), and bias. Both simulated and empirical latent trait distributions are used to sample vectors of true scores. MCATs were generally more efficient (in terms of test length) and more accurate (in terms of RMSE) than their UCAT counterparts. Absolute average bias was highest for variable-length UCATs with termination rule $SE < .387$. Test length of variable-length MCATs was on average 20% to 25% shorter than test length across separate UCATs. This study showed that there are clear advantages of using MCAT rather than UCAT in a setting typical for health measurement.

Keywords

multidimensional computerized adaptive testing, computerized adaptive testing, graded response model, item response theory, MCAT, M-CAT, UCAT, MIRT

[1]University of Groningen, Groningen, The Netherlands
[2]Centre for Educational Measurement (CEMO), University of Oslo, Oslo, Norway
[3]University of Twente, Enschede, The Netherlands
[4]VU University Medical Center, Amsterdam, The Netherlands
[5]Medisch Spectrum Twente, Enschede, The Netherlands

Corresponding Author:
Muirne C. S. Paap, Department of Special Needs Education and Youth Care, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Rozenstraat 38, 9712 TJ Groningen, The Netherlands.
Email: m.c.s.paap@rug.nl

Introduction

In the last decade, multidimensional computerized adaptive tests (MCATs; e.g., Segall, 1996, 2000) based on multidimensional item response theory (MIRT; e.g., Reckase, 2009) have become increasingly popular in applied settings, ranging from personnel selection to (mental) health measurement (Allen, Ni, & Haley, 2008; Gibbons et al., 2012; Makransky & Glas, 2013; Makransky, Mortensen, & Glas, 2013; Nikolaus et al., 2013; Nikolaus et al., 2015; Petersen et al., 2006). It is no surprise that MIRT models appeal to researchers in the fields of psychology and health measurement, as multifaceted constructs (such as psychopathology or quality of life) are the rule rather than the exception in these fields. Using MIRT allows one to concurrently estimate item parameters and a covariance structure among items/domains. Provided that the domains have nonzero correlations, items belonging to one domain can provide information about the latent trait scores on the other domains, either directly (if items load on more than one domain) or indirectly (through the correlations among the domains). This ''borrowing'' of information across domains allows for a more precise estimation of the latent trait values on the underlying domains compared with unidimensional IRT.

Studies comparing MCATs with unidimensional computerized adaptive tests (UCATs) can be roughly divided into those comparing an MCAT with one large UCAT (multidimensionality is ignored), and studies comparing an MCAT with performing a separate UCAT for each dimension. Here, we focus on the latter. The first studies explicitly comparing MCAT with separate UCATs focused on measuring ability (rather than, e.g., quality of life). These early studies showed that MCAT resulted in shorter tests (MCATs were 25%-33% shorter; Luecht, 1996; Segall, 1996), more accurate ability estimates (Li & Schafer, 2005), and a more balanced item exposure (Li & Schafer, 2005). Aforementioned studies varied in the type of model used: Whereas Segall (1996) and Li and Schafer (2005) used a three parameter logistic (3PL) model, Luecht (1996) used a multidimensional extension of the one parameter logistic model (OPLM). Other differences were whether content constraints were imposed, whether items were allowed to load onto multiple dimensions, whether the tests served a dual purpose, and the number of dimensions studied.

Ten years after the first ability MCAT studies were published, Haley, Ni, Ludlow, and Fragala-Pinkham (2006) bridged the gap with health measurement by comparing MCAT with UCAT for an existing pediatric health assessment tool. Using a multidimensional extension of the Rasch model, they found that using MCAT resulted in a 33% test reduction compared with using separate UCATs. Several authors—across different assessment settings—have shown that the added value associated with MCAT (in terms of test length and accuracy) increases as a function of the correlation among the domains (e.g., Frey & Seitz, 2010; Makransky & Glas, 2013; Makransky et al., 2013; W.-C. Wang & Chen, 2004). However, in a recent study focused on patient-reported outcomes, Bass, Morris, and Neapolitan (2015) reported a test reduction of only 11% when MCAT was compared with separate UCATs, in spite of the high correlation (r = .88) between the two dimensions under study. The authors tried to explain this unexpected finding by arguing that the studied items were highly informative, and therefore UCATs were already highly efficient, leaving little to no room for MCAT to further increase efficiency. However, Bass and colleagues used a maximum total test length of 20 items. As a result, up to 43% of their simulees did not reach the variable-length stopping rule. This may have had a substantial impact on their results.

Summarizing the available literature illustrating the feasibility and advantages of using MCAT in the field of health measurement, it can be observed that—although each of these studies made important contributions to the field—each of them had serious limitations when it comes to demonstrating the incremental value of MCAT over using separate UCATs. Some

articles focused only on dichotomous data (Allen et al., 2008; Haley et al., 2006), which is atypical for health measurement applications; other studies used very short item banks based on existing instruments not specifically developed for computerized adaptive testing (Michel et al., 2016; Petersen et al., 2006), while the study by Bass and colleagues had a relatively low maximum test length and was restricted to two dimensions (Bass et al., 2015). Nikolaus and colleagues advanced the field by studying the working mechanism of the MCAT they had developed to measure fatigue in patients with rheumatoid arthritis (Nikolaus et al., 2015), as well as patient acceptance (Nikolaus et al., 2014); they did not, however, explicitly compare MCAT with UCAT conditions.

As there is currently a marked increase in the use of CAT in psychological and health measurement, the main aim of the present study is to assess the benefits of using MCAT rather than separate UCATs for each domain with a setup typical for these fields. More specifically, data from an operational MCAT encompassing four correlated domains (developed to be used in a CAT) consisting of polytomous items are used. Given the results reported by Bass and colleagues, we were especially curious to see whether a more substantial reduction in test length would be observed when using MCAT (compared with separate UCATs). Similar to Bass and colleagues, the item parameters and covariance structure are estimated using empirical data and the multidimensional graded response model (GRM). We chose to study the effect of MCAT on test efficiency for both variable- and fixed-length conditions. In the variable-length condition, we opted for a high maximum test length, allowing us to truly study the effect of MCAT on a variable-length test.

Method

CATs are typically based on item banks calibrated with an IRT model. To facilitate MCAT, a multidimensional item bank calibrated with a MIRT model is required. Adams, Wilson, and Wang (1997) distinguished between two types of multidimensionality: within-item and between-item multidimensional models, which correspond to the ''simple'' and ''complex'' structures in factor analysis, respectively (W.-C. Wang & Chen, 2004). In this study, we focus on between-item multidimensionality, meaning that each item relates to one subdimension only; multidimensionality is expressed through the correlations among the latent dimensions (these are estimated jointly with the item parameters and latent trait values). The multidimensional GRM is used to obtain estimates of the item parameters and estimates of the covariance structure. The probability of a response in category j in item i with m total response categories, $P(X_{ij} = 1|\boldsymbol{\theta})$, is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi(\boldsymbol{\alpha}'\boldsymbol{\theta} - \beta_{i1}) & if \quad j = 0, \\ \Psi\left(\boldsymbol{\alpha}'\boldsymbol{\theta} - \beta_{ij}\right) - \Psi\left(\boldsymbol{\alpha}'\boldsymbol{\theta} - \beta_{i(j+1)}\right) & if \quad 0 < j < m, \\ \Psi(\boldsymbol{\alpha}'\boldsymbol{\theta} - \beta_{im}) & if \quad j = m, \end{cases}$$

where $\Psi(x)$ is the logistic function,

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)},$$

and $\boldsymbol{\alpha}'\boldsymbol{\theta}$ denotes the dot product of the vector of discrimination parameters and latent traits. To ensure that the probabilities are always positive, response categories must be sorted by difficulty, that is, $\beta_{i(j+1)} > \beta_{ij}$ for $0 < j < m$. The following parameters are calculated for each item: one discrimination parameter (denoted α) and a number of threshold parameters (denoted β_{ij}); the number of threshold parameters equals the number of categories minus 1.

Table 1. Covariance Matrix of Latent Traits Φ and Number of Items Per Domain.

	Fatigue	COPD-SIB	Physical function	Social roles
Fatigue	1	0.77	0.77	0.87
COPD-SIB		1	0.76	0.77
Physical function			1	0.84
Social roles				1
No. items per domain	50	46	63	35

Note. COPD-SIB = chronic obstructive pulmonary disease–specific item bank; social roles = ability to participate in social roles and activities.

Simulation Study

We start this section by briefly introducing the simulation design, after which we will illustrate the different components of the simulation design in more detail. The simulation study is based on an empirical multidimensional item bank that was developed to support MCAT, containing 194 items from four domains (subdimensions). The number of items per domain is listed in Table 1. The main design factor in the simulation is CAT type: MCAT versus UCAT. Two types of termination rules are also considered: fixed-length and variable-length/fixed precision. CAT performance is evaluated using three outcome variables: total length across domains, bias, and root mean square error (RMSE). Bias and RMSE are calculated per domain. Bias is here defined as the mean difference between CAT-based estimates $\hat{\theta}_{CAT}$ and the true θ value θ_{TRUE} (or complete item score pattern estimate θ_{COMPL}).[1] It expresses the degree to which the CAT estimates are systematically different from the true θ values (''accuracy''):

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_{CAT} - \theta_{TRUE}.$$

RMSE is measured as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(\hat{\theta}_{CAT} - \theta_{TRUE}\right)^2}.$$

Three sets of data (response patterns) are used to evaluate CAT performance: two synthetic datasets and one empirical. All responses for the synthetic datasets are generated based on the multidimensional GRM. For the empirical dataset, full-bank estimates are used as a proxy for true θ values. CAT simulations were run in R (R Development Core Team, 2012) using the package ShadowCAT[2] (Kroeze & de Vries, 2015).

The empirical item bank. The item bank consisted of the following domains: fatigue, physical function, and ability to participate in social roles and activities from the patient-reported outcome measurement information system (PROMIS; e.g., Cella et al., 2010; Terwee et al.; 2014); and a disease-specific set of items developed for patients with chronic obstructive pulmonary disease called the COPD-SIB (Paap, Lenferink, Herzog, Kroeze, & van der Palen, 2016). The calibration was performed using data from 795 patients with chronic obstructive pulmonary disease (COPD). The total number of items was distributed among three booklets each containing around 100 items, which were linked using 10 anchor items per domain (i.e., alternate form equating or common-item equating). Each booklet contained items pertaining to at least two

domains (see Supplement 1 for a graphical illustration of the booklet structure). Originally, all items in the item bank were scored on a 5-point scale ranging from 0 to 4. Following Paap et al. (2016), for 55 of 194 items, item response categories that showed low endorsement (fewer than 10 responses) were merged with adjacent categories. Higher scores were indicative of higher quality of life for all domains. The item bank was calibrated using the software package IRTPRO (Cai, Thissen, & du Toit, 2011). A multivariate normal distribution was assumed. To identify the model, the mean was set equal to 0, and variances were fixed to 1. The covariances were estimated freely. Note that in IRTPRO, an ''easiness'' parameter is estimated (denoted c) instead of difficulty; the β_{ij} parameter described above equals the negative value of IRTPRO's c parameter.

Experimental design factors. There are two main design factors: CAT type (two levels) and termination rule (five levels). In the MCAT conditions, the covariance matrix $\mathbf{\Phi}$ estimated using the multidimensional GRM is used as a prior. In the UCAT conditions, a separate UCAT is run for each domain with a $N(0,1)$ prior. The termination rule is either fixed-length ($k = 12$, $k = 20$, $k =$ 32 items) or variable-length/fixed precision (termination when $SE(\theta) < .387$, $SE(\theta) < .316$). For the MCAT conditions, the fixed-length k is specified *across domains*. For the UCAT conditions, the fixed-length is specified *per domain*: $k/4$. For variable-length conditions, item selection for a particular dimension is terminated when the *SE* threshold is reached for that particular dimension. Test developers and content experts who are not very familiar with IRT tend to favor reporting/interpreting the conditional reliability index—which is in fact a standardized *SE*—over information or *SE*. In this study, the *SE* termination rules were chosen such that they—at least in the unidimensional case and using a standard normal metric for θ—roughly correspond to reliability values of .85 and .90; these values are widely used in health measurement. See Raju, Price, Oshima, and Nering (2007) for more information on the topic of conditional reliability, including relevant equations.

CAT algorithm. The CAT algorithm used is illustrated in Figure 1. There are different starting rules to initialize a CAT (i.e., to obtain the initial θ value); two commonly used starting rules are to set θ at 0, or to administer a number of items to obtain an initial estimate. In this study, we chose the latter approach and administered a random item from each domain before initializing the CAT. Maximum a posteriori (MAP) estimates were used in all conditions in this study, as it allows prior knowledge to be introduced into the estimation process and has been shown to require fewer calculations than other Bayesian estimators such as the expected a posteriori estimator (Segall, 2000). Simulations were performed in a fashion largely identical to Segall's (1996) approach: The determinant of the posterior Fisher Information matrix was used as the objective function for item selection. Following this item-selection strategy, items are selected that provide the largest decrease of the size of the posterior credibility region (Segall, 2000). C. Wang, Chang, and Boughton (2013) referred to this item-selection method as the ''D-rule,''Diao and Reckase (2009) called it ''Bayesian Volume Decrease,'' and Yao (2013) simply abbreviated it as ''Volume'' or ''Vm.'' However, in the variable-length conditions item selection for a dimension was terminated when the *SE* for that particular dimension had reached the *SE* threshold. This prevented the CAT from selecting items that were technically superior but no longer needed to fulfill the stopping criteria. The maximum test length was set to 100 (to prevent the algorithm continuing to bank depletion in situations where ''high quality'' items were no longer available). For the UCAT conditions, separate CATs were run for each dimension; the starting, item-selection, and stopping procedures were largely equivalent to those used in the MCATs but adapted to a unidimensional setting, the only exception being how the fixed test length was set: across domains or by domain (see ''Experimental Design Factors'' section).

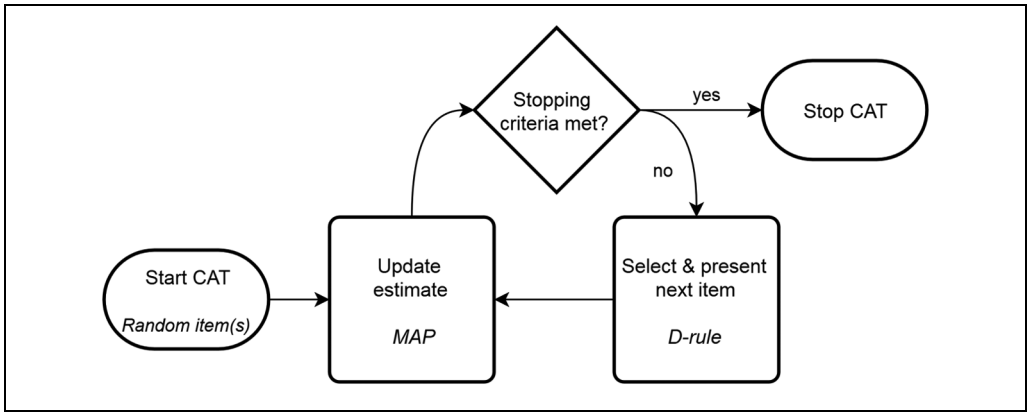

Figure 1. CAT flowchart.
Note. CAT = computerized adaptive test; MAP = maximum a posteriori.

Data generation. Three sets of data (response patterns) were used to evaluate CAT performance. Dataset 1 consists of synthetic response data, generated based on 21,000 vectors of discrete fixed latent trait points across the domains: 1,000 for every increment of .2 on the multidimensional θ scale between values −2 and 2. To handle the computational complexity that results from the four-dimensional ability space, the θ values considered for the four domains were set to be equal. In other words, a straight line is drawn through a four-dimensional space, where $\theta^{(1)} = \theta^{(2)} = \theta^{(3)} = \theta^{(4)}$. These vectors were chosen to explicitly evaluate the impact of the experimental design factors for more extreme trait scores (which are relatively rare when a multivariate normal distribution is assumed). Dataset 2 consists of synthetic response data, generated based on 1,000 vectors of true θ values sampled from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Phi}$. Dataset 3 consists of the empirical response patterns that were used to estimate the multidimensional GRM (missing data due to the booklet structure was imputed using the multidimensional GRM). The latter two datasets allow us to evaluate the impact of the experimental design factors under realistic circumstances. A more detailed description of the data-generation procedure follows below.

All responses were generated based on the multidimensional GRM. Responses were generated by comparing a draw u from a $U(0, 1)$ distribution to the cumulative probabilities of endorsing a response category j, $P_{cum}(x_j) = \sum_{q=0}^{j} P(x_q | \ldots)$. The generated response is equal to the lowest response category where the cumulative probability is greater than the random draw:

$$u \sim U(0,1)$$
$$x = \min_j P_{cum}(x_j) > u.$$

Given that the cumulative probability of the highest response category is always 1, this guarantees an answer. ShadowCAT uses R's built-in uniform distribution for this purpose. For the second dataset, simulees were drawn from a $N(\mathbf{0}, \mathbf{\Phi})$ distribution, where the off-diagonal elements in the population variance–covariance matrix $\mathbf{\Phi}$ were specified from the empirical correlations among the four domains. Given the booklet design, not all data were available for the empirical sample. Rather than constrain the CATs to use only available responses, we decided to impute the missing responses based on available data. This was done in two steps: First, for all respondents, a θ vector was estimated using a MAP estimator, with a mean vector of $\mathbf{0}$, and

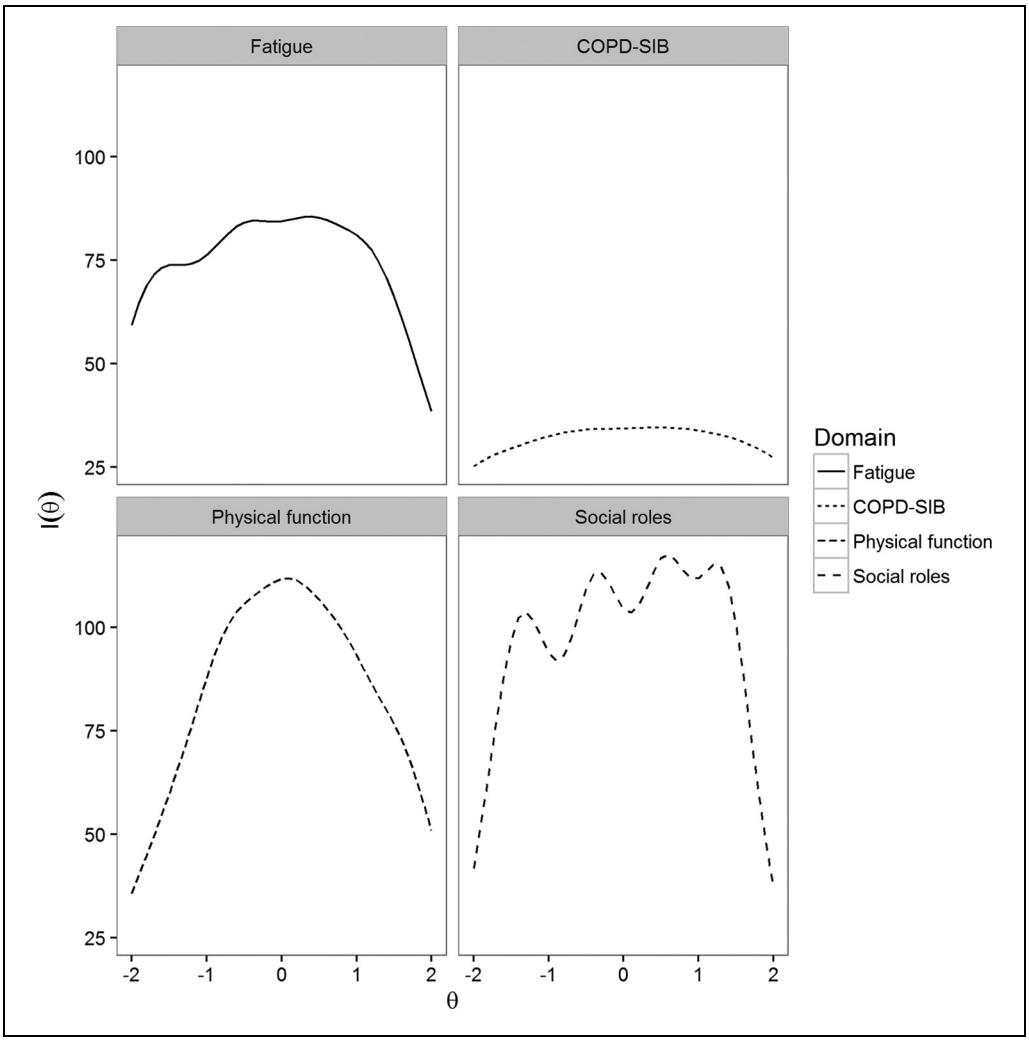

Figure 2. Bank information functions for all four domains.
Note. COPD-SIB = chronic obstructive pulmonary disease–specific item bank.

prior covariance equal to the population covariance. Thus, a multivariate structure was assumed in imputation for both unidimensional and multidimensional simulations. Second, missing data were imputed using the same response generation method that was used for simulated respondents.

Results

Multidimensional Calibration

Figure 2 presents the bank information functions for each domain. It can be seen that the entire range of latent trait values is covered, and information is fairly high for all domains, especially for the PROMIS domains; this is a common finding for clinical/health-related tests, where

Table 2. Average Total Test Length for Variable-Length CATs and All Three Datasets (Columns).

		Dataset 1					Dataset 2	Dataset 3
CAT type	Termination rule, *SE*	−2	−1	0	1	2	Normal simulated	Empirical sample
Total (across domains)								
MCAT	<0.387	12.9	9.6	9.1	8.7	11.8	9.5	9.6
	<0.316	20.4	14.1	13.2	13.0	18.8	14.4	14.8
UCAT	<0.387	16.7	12.7	11.7	11.3	14.7	12.3	12.4
	<0.316	26.1	17.8	16.6	16.2	23.9	18.5	18.8
Fatigue								
MCAT	<0.387	2.4	2.0	2.0	2.0	2.7	2.1	2.1
	<0.316	3.5	3.0	2.9	3.0	4.2	3.1	3.2
UCAT	<0.387	3.1	2.9	2.8	2.9	4.5	3.0	3.0
	<0.316	4.3	4.0	3.8	3.8	5.7	4.2	4.3
COPD-SIB								
MCAT	<0.387	4.2	3.7	3.5	3.2	3.7	3.5	3.5
	<0.316	7.5	6.0	5.7	5.5	6.5	5.9	6.0
UCAT	<0.387	6.0	4.9	4.6	4.4	5.0	4.7	4.7
	<0.316	9.7	7.4	6.9	6.8	8.0	7.3	7.4
Physical function								
MCAT	<0.387	3.9	2.2	2.0	2.0	3.1	2.2	2.2
	<0.316	5.7	3.0	2.6	2.6	4.3	3.0	3.1
UCAT	<0.387	4.6	2.7	2.2	2.1	2.8	2.5	2.5
	<0.316	6.5	3.5	3.1	3.0	4.3	3.6	3.6
Ability to participate in social roles and activities								
MCAT	<0.387	2.3	1.7	1.6	1.6	2.2	1.7	1.8
	<0.316	3.7	2.1	2.0	2.0	3.9	2.4	2.5
UCAT	<0.387	2.9	2.3	2.1	2.0	2.4	2.2	2.2
	<0.316	5.7	3.0	2.8	2.6	5.9	3.4	3.6

Note. CATs = computerized adaptive tests; MCAT = multidimensional CAT; UCAT = unidimensional CAT; COPD-SIB = chronic obstructive pulmonary disease-specific item bank.

constructs are often conceptually narrow and thus have high discrimination parameters (Reise & Waller, 2009). Item exposure rates for the MCAT and UCAT conditions can be found in Supplement 1.

CAT Simulation Studies

Comparison of the MCAT and UCAT conditions. The average test lengths (averaged over domains) for the various CAT conditions can be found in Table 2. It shows that MCAT outperforms UCAT in terms of test length for both *SE*-based termination rules and all datasets. MCATs were on average 20% to 25% shorter than UCATs for Dataset 1, 22% to 23% shorter for Dataset 2 (simulated normal distribution), and 21% to 23% shorter for Dataset 3 (empirical sample).

Bias and RMSE for all CAT conditions based on Dataset 2 and Dataset 3 can be found in Tables 3 and 4, respectively. Average bias was low for most MCAT conditions, especially for the fixed-length tests. Average RMSE across termination rules and domains was somewhat lower for MCAT (Dataset 2: 0.283 and 0.305 for MCAT and UCAT, respectively; Dataset 3: 0.259 and 0.292 for MCAT and UCAT, respectively). Thus, on average, MCATs were shorter than UCATs while resulting in similar or better RMSE. Figures 3 and 4 show bias and RMSE for the two variable-length termination rules based on Dataset 1. Figure 3 shows MCATs and UCATs performed equally well in terms of bias for average θ values, but MCATs always

Table 3. Average Bias for the Four Domains (Dataset 2 and Dataset 3).

		Dataset 2 (normal simulated)				Dataset 3 (empirical sample)			
CAT type	Termination rule	FAT	CSIB	PHYS	SOC	FAT	CSIB	PHYS	SOC
MCAT	$SE < .387$	−0.006	0.023	−0.007	−0.008	−0.026	−0.009	−0.017	−0.017
	$SE < .316$	0.015	−0.004	0.006	0.017	−0.002	−0.002	−0.015	−0.012
UCAT	$SE < .387$	0	0.021	0.006	0.029	−0.010	−0.019	−0.014	−0.021
	$SE < .316$	−0.014	0.007	0	−0.009	0.007	−0.014	−0.002	−0.021
MCAT	Fixed 12	0.006	0.007	0.003	0.003	−0.011	−0.001	−0.006	−0.013
	Fixed 20	0.001	−0.001	0.003	−0.005	−0.004	−0.005	−0.007	−0.006
	Fixed 32	0.002	0.004	0.01	0.002	0.005	−0.006	0.003	−0.004
UCAT	Fixed 12	−0.005	0.003	0.016	0.003	−0.014	−0.008	−0.013	−0.024
	Fixed 20	0.013	−0.006	0.002	−0.002	−0.010	−0.020	−0.008	−0.011
	Fixed 32	−0.001	0	−0.003	−0.008	0.001	−0.012	−0.001	−0.012

Note. CAT = computerized adaptive test; FAT = fatigue; CSIB = chronic obstructive pulmonary disease-specific item bank; PHYS = physical function; SOC = ability to participate in social roles and activities; MCAT = multidimensional CAT; UCAT = unidimensional CAT.

Table 4. Average RMSE for the Four Domains (Dataset 2 and Dataset 3).

		Dataset 2 (normal simulated)				Dataset 3 (empirical sample)			
CAT type	Termination rule	FAT	CSIB	PHYS	SOC	FAT	CSIB	PHYS	SOC
MCAT	$SE < .387$	0.352	0.367	0.340	0.333	0.344	0.352	0.298	0.280
	$SE < .316$	0.298	0.317	0.306	0.281	0.295	0.292	0.257	0.249
UCAT	$SE < .387$	0.349	0.377	0.368	0.347	0.355	0.384	0.342	0.351
	$SE < .316$	0.311	0.305	0.300	0.29	0.285	0.309	0.278	0.259
MCAT	Fixed 12	0.294	0.391	0.284	0.254	0.279	0.381	0.253	0.210
	Fixed 20	0.238	0.330	0.229	0.200	0.221	0.325	0.208	0.177
	Fixed 32	0.200	0.276	0.200	0.172	0.177	0.274	0.181	0.145
UCAT	Fixed 12	0.332	0.469	0.340	0.301	0.354	0.461	0.316	0.272
	Fixed 20	0.284	0.357	0.258	0.229	0.261	0.367	0.243	0.221
	Fixed 32	0.216	0.284	0.207	0.184	0.204	0.292	0.194	0.173

Note. RMSE = root mean square error; CAT = computerized adaptive test; FAT = fatigue; CSIB = chronic obstructive pulmonary disease-specific item bank; PHYS = physical function; SOC = ability to participate in social roles and activities; MCAT = multidimensional CAT; UCAT = unidimensional CAT.

resulted in lower absolute bias than UCATs for more extreme θ values. Moreover, Figure 4 shows that MCATs resulted in lower RMSE values than UCATs for all conditions, domains, and θ values. Figures 5 and 6 show bias and RMSE for the three fixed-length termination rules based on Dataset 1; these figures illustrate that MCATs generally performed equally well or better in terms of bias and RMSE for all conditions and domains. The benefits of using MCAT for fixed-length tests are most pronounced for the higher θ values and the 12-item termination rule.

Comparison of the termination rules: Variable-length versus fixed-length. Table 3 shows that fixed-length MCATs were associated with the lowest absolute average bias. Absolute average bias was highest for variable-length UCATs with termination rule $SE < .387$. When looking at Table 4, a striking domain-related pattern emerges for RMSE. If we take Dataset 2 as an example, we see that—for all PROMIS domains—the fixed-length tests resulted in the lowest RMSE values when MCAT was used. When UCATs were used, the lowest RMSE values were found

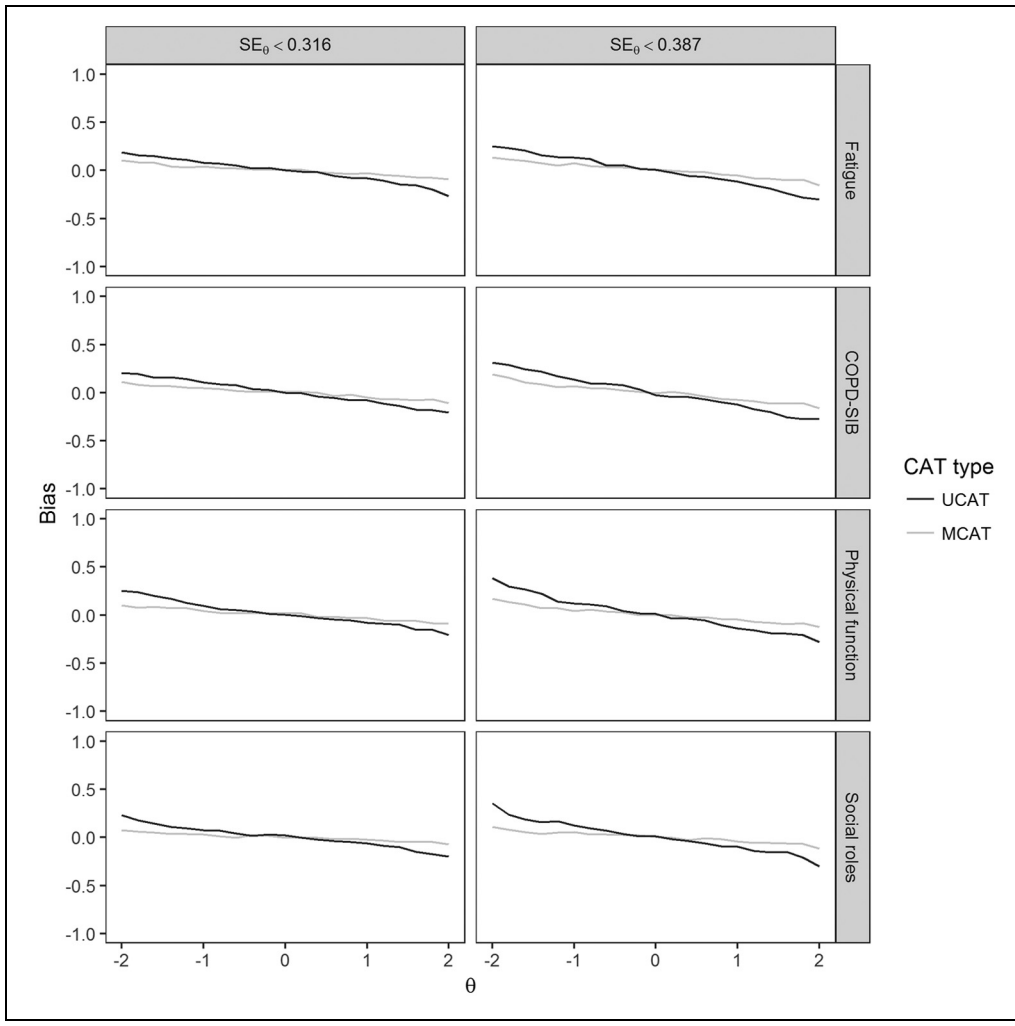

Figure 3. Bias as a function of the true θ value on a given domain using Dataset 1.
Note. It is shown for variable-length CATs. CAT = computerized adaptive test; COPD-SIB = chronic obstructive pulmonary disease-specific item bank; UCAT = unidimensional CAT; MCAT = multidimensional CAT.

for 32-item fixed-length tests and the $SE < .316$ termination rule. For the COPD-SIB, the pattern that emerges is the same for MCAT and UCAT: The lowest RMSE was found for 32-item fixed-length tests, followed by the $SE < .316$ termination rule, and 20-item fixed-length tests. Inspection of Figures 3 to 6 shows that bias and RMSE differ as a function of the latent trait. RMSE is clearly highest for the following combination of conditions: UCAT, 12-item fixed-length tests, more extreme θ values, COPD-SIB domain.

Discussion

This study shows a clear benefit of using MCAT rather than separate UCATs: MCATs were generally more efficient (in terms of test length) and more accurate (in terms of bias and RMSE) than their UCAT counterparts. The gains in accuracy were largest for fixed-length tests,

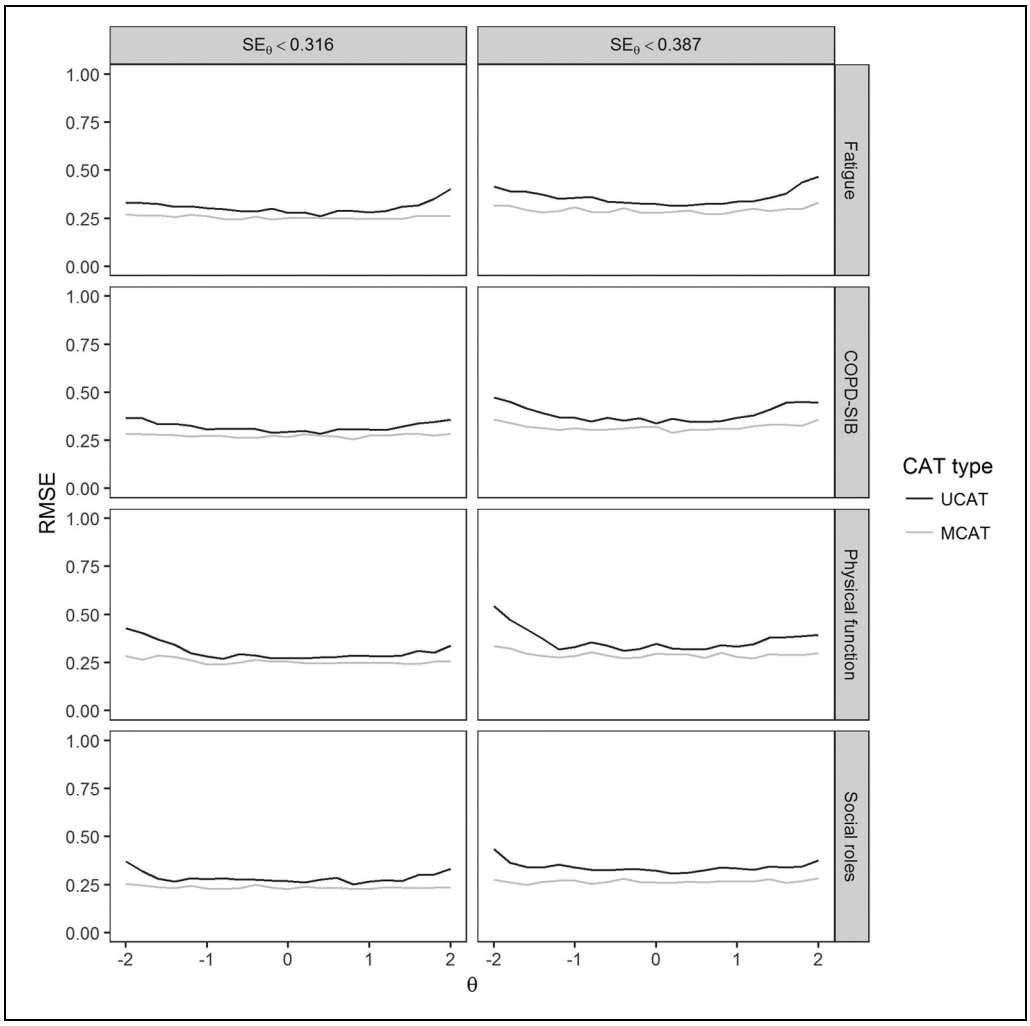

Figure 4. RMSE as a function of the true θ value on a given domain using Dataset 1.
Note. It is shown for variable-length CATs. RMSE = root mean square error; CAT = computerized adaptive test; COPD-SIB = chronic obstructive pulmonary disease-specific item bank; UCAT = unidimensional CAT; MCAT = multidimensional CAT.

and most pronounced for higher θ values, short tests (k = 12), and the domain with the lowest discrimination parameters.

We compared MCAT with UCAT for two types of termination rules: fixed-length and variable-length. MCAT outperformed UCAT for both types of rules. The findings of the present study suggest that, overall, fixed-length tests perform well in terms of latent trait estimation accuracy when using MCAT based on highly correlated dimensions containing many good quality items (high discrimination parameters). For dimensions with lower discrimination parameters, variable-length MCATs may be more suitable, especially for more extreme θ values as in such instances, fixed-length CATs may be too short to provide accurate measurement. In this study, we used the D-rule/Bayesian Volume Decrease item-selection rule. As we were interested in optimizing measurement efficiency for each of the dimensions, items were selected

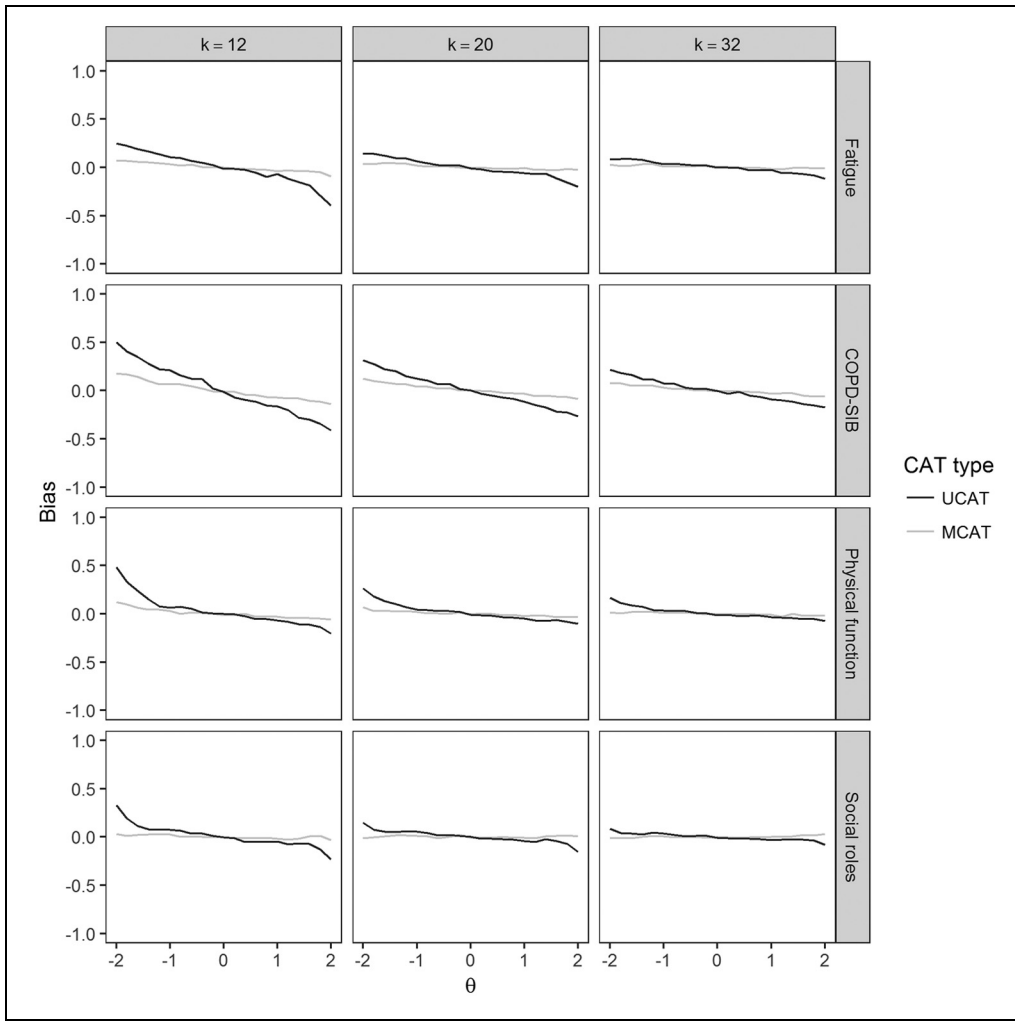

Figure 5. Bias as a function of the true θ value on a given domain using Dataset 1 for three fixed-length termination rules.

Note. *k* = maximum number at which test is terminated; CAT = computerized adaptive test; COPD-SIB = chronic obstructive pulmonary disease-specific item bank; UCAT = unidimensional CAT; MCAT = multidimensional CAT.

only from those dimensions whose *SE* threshold had not been reached yet (variable-length MCAT conditions). More research is needed to investigate whether our findings can be generalized to other item-selection rule and stopping rule combinations. Relevant item-selection rules could include the E-rule (minimum eigenvalue rule), T-rule (maximum trace rule), and K-rule (maximum Kullback–Leibler divergence rule) described by C. Wang et al. (2013), in combination with the *SE* and the predicted standard error reduction (PSER) stopping rules (e.g., Yao, 2013).

As has been pointed out previously, item selection and parameter estimation under an MCAT are rather complex, and implementing MCAT methodology may require a larger investment in terms of resources compared with a UCAT. However, the present study indicates that it may well be worth it: A lot can be gained by incorporating information regarding the

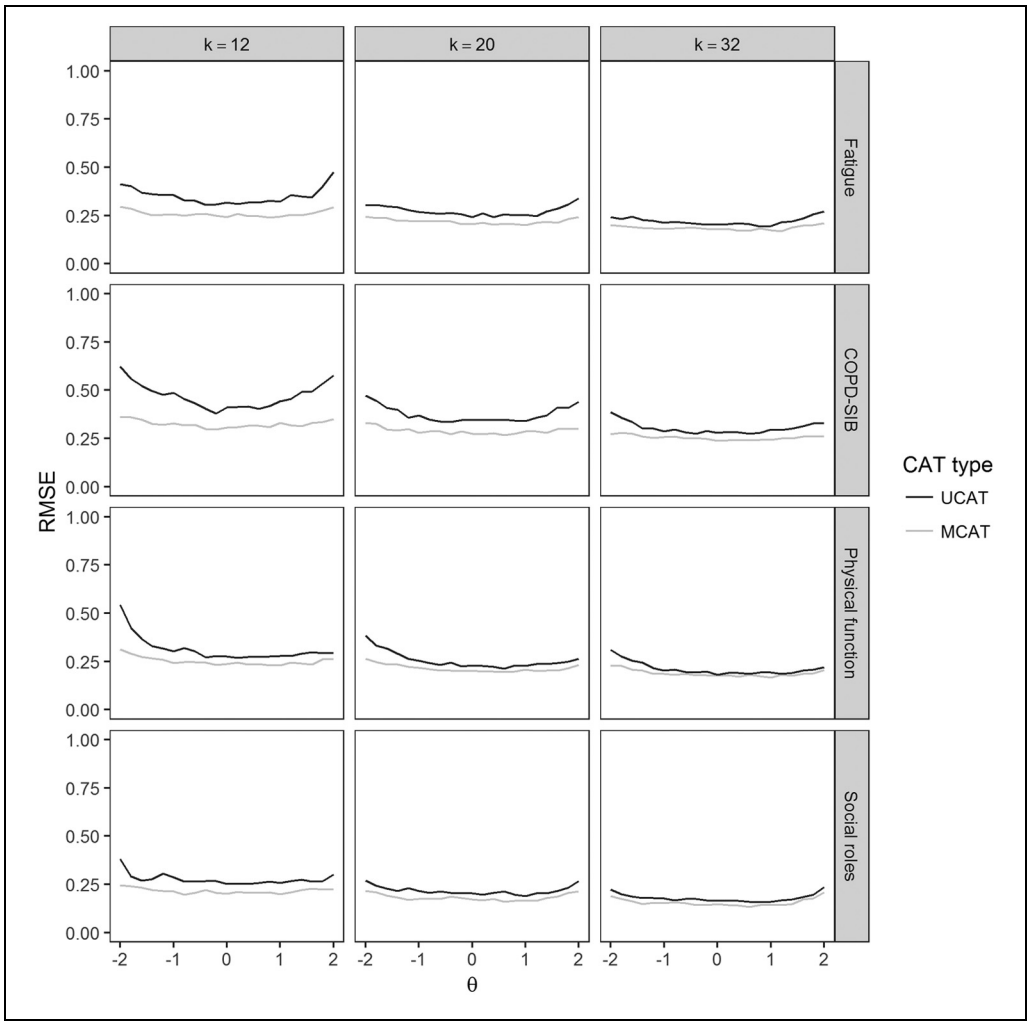

Figure 6. RMSE as a function of the true θ value on a given domain using Dataset 1 for three fixed-length termination rules.

Note. *k* = maximum number at which test is terminated; RMSE = root mean square error; CAT = computerized adaptive test; COPD-SIB = chronic obstructive pulmonary disease-specific item bank; UCAT = unidimensional CAT; MCAT = multidimensional CAT.

covariance structure among domains into the item-selection process in a CAT, also for polytomously scored items. Polytomous items are generally richer in information (they cover a wider range of θ values) than dichotomously scored items, which could theoretically lead to smaller benefits of ''borrowing'' additional information across domains compared with a dichotomous item bank scenario. It would be interesting to compare dichotomous and polytomous item banks directly in a future study when comparing UCAT with MCAT.

In the current study, we used a multidimensional item bank with four domains, calibrated using empirical polytomous data, which allowed us to show the benefits of MCAT under realistic testing conditions typical for patient-reported outcomes. CAT is gaining momentum in

psychological and health measurement, and therefore our findings may be especially valuable to test developers and administrators in these fields.

Acknowledgments

We wish to thank the staff of the following clinics for their assistance with data collection: Medisch Spectrum Twente, Sint Lucas Andreas Ziekenhuis, CIRO Center of Expertise for Chronic Organ Failure, Martini Ziekenhuis Groningen, Scheperziekenhuis, Sint Franciscus Gasthuis, Canisius-Wilhelmina Ziekenhuis, VU University Medical Center, Twentse Huisartsen Onderneming Oost Nederland, Gelre Ziekenhuizen, and the University Medical Center Groningen, as well as all participating physiotherapists. Special thanks to Siebrig Schokker for her invaluable contribution to the data collection. We thank all patients who participated in the study.

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.

Authors' Note

Opinions reflect those of the authors and do not necessarily reflect those of the funding agency.

Funding

The author(s) disclosed receipt of the following financial support for the research described in this article: This study was supported by a grant from Lung Foundation Netherlands (grant #3.4.11.004).

Supplemental Material

Supplementary material is available for this article online.

Notes

1. For the computerized adaptive test simulations based on empirical data, no ''true''θ is available; in this scenario, the maximum a posteriori (MAP) estimate of θ based on all available item responses (denoted θ_{COMPL}) is used as a substitute for the true θ value.
2. ShadowCAT facilitates simulation and live administration of CATs. Options include four generalized multidimensional models (three-parameter logistic model, generalized partial credit model, graded response model, sequential model), three estimators (maximum likelihood, maximum a posteriori, expected a posteriori), Fisher Information, and Kullback–Leibler distance-based item selection, constraints, content balancing and exposure control with Shadow Testing.

References

Adams, R. J., Wilson, M., & Wang W-c. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Allen, D. D., Ni, P., & Haley, S. M. (2008). Efficiency and sensitivity of multidimensional computerized adaptive testing of pediatric physical functioning. *Disability and Rehabilitation*, *30*, 479-484.

Bass, M., Morris, S., & Neapolitan, R. (2015). Utilizing multidimensional computer adaptive testing to mitigate burden with patient reported outcomes. *AMIA Annual Symposium Proceedings*, *2015*, 320-328.

Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO 2.1 [Computer software]. Lincolnwood, IL: Scientific Software International.

Cella, D. F., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Choi, S. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology*, *63*, 1179-1194.

Diao, Q., & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [08/08/2016] from http://www.iacat.org/content/comparison-ability-estimation-and-item-selection-methods-multidimensional-computerized

Frey, A., & Seitz, N.-N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz [Multidimensional adaptive diagnostics of competences: Results on measurement efficiency]. *Zeitschrift Für Pädagogik, Beiheft*, *56*, 40-51.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., . . . Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*, 1104-1112.

Haley, S. M., Ni, P., Ludlow, L. H., & Fragala-Pinkham, M. A. (2006). Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation*, *87*, 1223-1229.

Kroeze, K. A., & de Vries, R. (2015). *ShadowCAT: Multidimensional computer adaptive testing with the Shadow Testing routine*. Retrieved from https://github.com/Karel-Kroeze/ShadowCAT/

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement*, *29*, 3-25.

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

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*, 123-139.

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the NEO PI-R. *Assessment*, *20*(1), 3-13.

Michel, P., Baumstarck, K., Ghattas, B., Pelletier, J., Loundou, A., Boucekine, M., . . . Boyer, L. (2016). A Multidimensional Computerized Adaptive Short-Form Quality of Life Questionnaire developed and validated for multiple sclerosis: The MusiQoL-MCAT. *Medicine*, *95*(14), Article e3068.

Nikolaus, S., Bode, C., Taal, E., Oostveen, J. C., Glas, C. A. W., & van de Laar, M. A. F. J. (2013). Items and dimensions for the construction of a multidimensional computerized adaptive test to measure fatigue in patients with rheumatoid arthritis. *Journal of Clinical Epidemiology*, *66*, 1175-1183.

Nikolaus, S., Bode, C., Taal, E., Vonkeman, H. E., Glas, C. A. W., & van de Laar, M. A. F. J. (2014). Acceptance of new technology: A usability test of a Computerized Adaptive Test for fatigue in rheumatoid arthritis. *JMIR Human Factors*, *1*(1), Article e4.

Nikolaus, S., Bode, C., Taal, E., Vonkeman, H. E., Glas, C. A. W., & van de Laar, M. A. F. J. (2015). Working mechanism of a multidimensional computerized adaptive test for fatigue in rheumatoid arthritis. *Health Quallife Outcomes*, *13*, Article 23.

Paap, M. C. S., Lenferink, L. I. M., Herzog, N., Kroeze, K. A., & van der Palen, J. (2016). The COPD-SIB: A newly developed disease-specific item bank to measure health-related quality of life in patients with chronic obstructive pulmonary disease. *Health and Quality of Life Outcomes*, *14*, Article 97.

Petersen, M. A., Groenvold, M., Aaronson, N., Fayers, P., Sprangers, M., & Bjorner, J. B. (2006). Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluations. *Quality of Life Research*, *15*, 315-329.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*, 169-180.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27-48.

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

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53-74). Dordrecht, the Netherlands: Kluwer Academic.

Terwee, C. B., Roorda, L. D., de Vet, H. C., Dekker, J., Westhovens, R., van Leeuwen, J.,. . . Boers, M. (2014). Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). *Quality of Life Research*, *23*, 1733-1741.

Wang, C., Chang, H.-H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*, 99-122.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 295-316.

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