Qual Life Res
DOI 10.1007/s11136-017-1624-3

Item usage in a multidimensional computerized adaptive test (MCAT) measuring health-related quality of life

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

Accepted: 13 June 2017

Abstract
Purpose Examining item usage is an important step in evaluating the performance of a computerized adaptive test (CAT). We study item usage for a newly developed multidimensional CAT which draws items from three PROMIS domains, as well as a disease-specific one.
Methods The multidimensional item bank used in the current study contained 194 items from four domains: the PROMIS domains fatigue, physical function, and ability to participate in social roles and activities, and a disease-specific domain (the COPD-SIB). The item bank was calibrated using the multidimensional graded response model and data of 795 patients with chronic obstructive pulmonary disease. To evaluate the item usage rates of all individual items in our item bank, CAT simulations were performed on responses generated based on a multivariate uniform distribution. The outcome variables included active bank size and item overuse (usage rate larger than the expected item usage rate).
Results For average θ-values, the overall active bank size was 9–10%; this number quickly increased as θ-values became more extreme. For values of -2 and $+2$, the overall active bank size equaled 39–40%. There was 78% overlap between overused items and active bank size for average θ-values. For more extreme θ-values, the overused items made up a much smaller part of the active bank size: here the overlap was only 35%.
Conclusions Our results strengthen the claim that relatively short item banks may suffice when using polytomous items (and no content constraints/exposure control mechanisms), especially when using MCAT.

Keywords Item exposure · HRQL · IRT · Item response theory · MCAT · CAT · MAT · Computerized adaptive test

Introduction

In the last decade, computerized adaptive tests (CATs) [1] based on item response theory (IRT) [2] have become increasingly popular in health measurement. A CAT can be seen as a questionnaire that is tailored to the test-taker on the fly: it continuously updates the estimate(s) of the position on the construct of interest (*latent trait*) based on answers given by the test-taker to the questions (items) posed. The underlying algorithm then selects the item that is most informative at that particular moment, given the current estimate of the latent trait value. It is clear why

Electronic supplementary material The online version of this article (doi:10.1007/s11136-017-1624-3) contains supplementary material, which is available to authorized users.

✉ Muirne C. S. Paap
m.c.s.paap@rug.nl

1 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

2 Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Oslo, Norway

3 Department of Research Methodology, Measurement, and Data-Analysis, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

4 Department of Epidemiology and Biostatistics and the EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands

5 Medical School Twente, Medisch Spectrum Twente, Enschede, The Netherlands

Published online: 23 June 2017

CATs appeal to healthcare professionals (HCPs): by selecting only those items that contribute most to the reliable measurement of a patient's latent trait value, measurement efficiency is increased, which results in a substantial decrease in response burden [3]. Furthermore, CAT estimates can be used to generate automatic reports instantly, providing the HCP with all necessary information (latent trait estimate, standard error, norms, and graphic display) to facilitate communication with the patient. These properties make CATs excellent candidates for monitoring patients' physical and mental health routinely, be it on a monthly or daily basis.

CATs draw their items from item banks: large collections of items that have been calibrated with an IRT model using a large sample representative of the target population. The quality of the CAT and the latent trait estimate it generates depend to a large degree on the quality of the item bank. A psychometrically sound item bank contains items with location parameters that cover the whole range of relevant latent trait values, while having adequate to high discrimination parameters. A CAT drawing items from such an item bank will result in efficient measurement for all patients (irrespective of their latent trait score). Most CATs currently used for health measurement are based on item banks that were calibrated using unidimensional IRT models (e.g., [4–7]). Although less frequently used, multidimensional IRT models are available as well, and can be used to support multidimensional CAT (MCAT) (e.g., [8–10]). It has been shown that test length can be further reduced by taking the correlation among constructs into account during item selection and latent trait estimation, while maintaining adequate levels of measurement precision [11, 12]. Perhaps equally important, patients often experience quality-of-life (QoL) domains as interdependent; taking this into account allows a closer alignment between psychometric modeling and patient perspective.

Since health-related quality of life (HRQL) has taken a central role in the evaluation of treatment interventions in patients with chronic obstructive pulmonary disease (COPD), we recently developed a multidimensional CAT (MCAT) to measure HRQL in patients with chronic obstructive pulmonary disease (COPD) [13]. Following the steps outlined by Paap et al. [14], we first established which domains of HRQL are most important to patients with COPD, using relevant literature (articles and existing questionnaires), as well as interviews with patients and HCPs [14, 15]. Based on these findings, three generic domains/item banks from the PROMIS (Patient-Reported Outcomes Measurement Information System) framework were selected (fatigue, physical functioning, and ability to participate in social roles and activities) and a new COPD-specific domain/item bank (COPD-SIB) was developed [16]. This approach ensures comparability with other patient groups (generic domains), while providing additional sensitivity for measuring change within the specific patient group (disease-specific domain). In this paper, we aim to evaluate an important performance measure for our CAT: *item usage*.

Due to the adaptive nature of a CAT, it can be expected that certain items are used more frequently than others. Successive items are typically chosen to optimize an objective function [17], such as the Fisher information function.[1] Highly discriminating items, polytomous items covering a wide range of the latent trait (denoted θ), and items targeting average θ-values have a higher chance of being selected, all else being equal. If items are selected more frequently than could be expected based on chance or a predefined threshold, these items are typically referred to as being overexposed. Conversely, items selected less frequently than could be expected are referred to as underexposed. The terms *item exposure* and *item usage* seem to be used interchangeably in the literature. In the context of educational testing, item overexposure is seen as a threat to test security (examinees may be able to remember and share items with others) and receives a lot of attention in the literature (see, e.g., [18, 20, 21]); in health measurement, items do not need to be kept secret and therefore item exposure has received less attention [22]. However, item usage is an important outcome measure in evaluating CAT and item bank performance. Variability in item usage rates indicates that the CAT is working as intended (if the items were selected at random, the item usage rate would be expected to be equal for all items). However, if a number of items are not used at all, or very rarely, the "real" (active) size of the item bank is smaller than it was designed to be. The main aim of the current study is to evaluate item usage for a newly developed MCAT which draws items from the PROMIS domains fatigue, physical function, and ability to participate in social roles and activities, as well as the COPD-SIB. We will report on both active bank size and item overuse/overexposure.

[1] In a unidimensional setting, Fisher information, which varies as a function of the latent trait value, is often used. When the item location parameter equals the latent trait value, Fisher information increases monotonically as the value of the discrimination parameter increases; therefore, the item selection rule based on Fisher information will select an item with a location close to the current latent trait estimate and a discrimination that is as large as possible (see, e.g., [18]). In MCAT, item information is no longer expressed by a single value; instead, item selection typically depends on the value of the determinant of the posterior information matrix (this value is computed and evaluated for each of the remaining items in the multidimensional item bank, and the item for which the value is largest is selected) [19].

Methods

Multidimensional item bank

Adams et al. [23] divide multidimensional IRT models into two subclasses: *within-item* and *between-item* multidimensional models. Within-item multidimensional models allow items to relate to more than one latent dimension. When between-item multidimensional models are used, the restriction is imposed that the items relate to one dimension only; multidimensionality is expressed through the correlations among the latent dimensions (these are estimated jointly with the item parameters and latent trait values). In this study, we chose to use a between-item multidimensional model, since such models are useful when multiple distinct latent dimensions are measured[2] and relatively high correlations are expected. The multidimensional item bank used in the current study contained 194 items from four domains: the PROMIS domains fatigue (example item: "To what degree did you have to push yourself to get things done because of your fatigue?"), physical function (example item: "Are you able to climb up five steps?"), and ability to participate in social roles and activities (example item: "I have trouble doing all of the activities with friends that are really important to me") [25, 26]; and the COPD-SIB (example item: "It frustrated me that I couldn't do everything I wanted to do anymore") [16]. The PROMIS ability to participate in social roles and activities item bank was used in its entirety (35 items). We included a subset of the other two PROMIS item banks: we selected 50 fatigue and 63 physical function items. Item selection was performed by JP who has ample experience with COPD patients and COPD research, and reviewed by an international colleague of JP's with comparable experience. The COPD-SIB contains 46 items: both newly written items, and (adapted versions of) items from the SGRQ-C, the Quality of Life for Respiratory Illness Questionnaire (QoL-RIQ), the COPD Assessment Test, the Maugeri Respiratory Failure Questionnaire Reduced Form (MRF26), and the VQ11 [27–30]. In our application, a higher latent trait score indicated better HRQL for all domains.

Test design

Multidimensional calibrations are not currently available for the PROMIS general population sample, and therefore the PROMIS calibrations cannot be used in the current study. In order to facilitate multidimensional calibration, our test design needed to be constructed in a way that would allow for item parameter estimation as well as estimation of the covariance structure among the domains. We used a booklet design, whereby the total number of items was distributed among three booklets each containing around 100 items. The booklets were linked using ten anchor items per domain (this type of linking is also known as *alternate form equating* or *common-item equating*). Each booklet contained items pertaining to at least two domains.

Calibration sample

The following inclusion criteria were used: a medical diagnosis of COPD; sufficient oral and written mastery of the Dutch language; and being able to complete a questionnaire. HCPs (pulmonologists, general practitioners, physiotherapists, and nurse practitioners) were recruited by JP, through his professional network. HCPs distributed the questionnaires accompanied by an information letter among COPD patients attending their clinics from October 2014 through December 2015. Of the 1500 printed booklets, 795 were returned by the end of December 2015. Our sample had a mean age of 67.2 years ($SD = 10.08$), and consisted of 52.7% men. More detailed patient characteristics are reported in Supplement 1.

Data preparation

All items in the item bank were scored on a 5-point Likert scale ranging from 0 to 4. In total, 10 different types of answer categories were used (depending on the domain and item formulation), for example, *without any difficulty, with a little difficulty, with some difficulty, with much difficulty, unable to do* or *never, rarely, sometimes, usually, always.* Twenty-eight percent of the items showed low endorsement (fewer than 10 responses) for one or more of its categories. Following Paap et al. [16], for 55 out of 194 items, item response categories that showed low endorsement (fewer than 10 responses) were merged with adjacent categories. Among these 55 items, 18 pertained to the fatigue domain, 23 to physical function, 2 to ability to participate in social roles and activities, and 12 to the COPD-SIB. For the majority of these items (51), the lowest two or highest two categories were collapsed. In the other cases, either the lowest or highest three categories were collapsed, or both the lowest two and the highest two. Note that items having different numbers of response categories due to merging does not constitute a problem for the IRT model used (multidimensional GRM).

[2] In interviews with healthcare professionals [24], the target group that was to use our CAT, the majority indicated that they were not interested in a global score, but instead favored separate scores for each dimension (data not shown).

Multidimensional IRT calibration

The multidimensional graded response model was used to obtain item parameter estimates 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

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

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, $\beta_{i(j+1)} > \beta_{ij}$ for $0 < j < m$.

Up to five parameters were calculated for each item i: one discrimination parameter (denoted α_i) and several β_{ij} parameters; the number of β_{ij} parameters equals the number of categories minus one. The β_{ij} parameter is related to the difficulty with which a respondent will reach the jth step of each item. Note that in unidimensional IRT, two types of parametrization can be used for x: $\alpha(\theta - \beta)$ or $\alpha\theta - \beta$. In multidimensional IRT, $\boldsymbol{\alpha}'$ is a vector containing an α value for each dimension; here, only the $\alpha\theta - \beta$ parametrization can be used. Some software packages, such as IRTPRO, calculate "easiness" rather than "difficulty" parameters. In IRTPRO, this parameter is denoted as c. The β_{ij} parameter described above equals the negative value of the c-parameter. The estimates of the item parameters and covariance structure were obtained using the software package IRTPRO [31].

A multivariate normal distribution was assumed for the four latent traits, with variances fixed to 1 and the covariances being estimated freely. The estimated correlation matrix among the four domains Φ equalled

$$\begin{bmatrix} 1 & 0.77 & 0.87 & 0.77 \\ 0.77 & 1 & 0.84 & 0.76 \\ 0.87 & 0.84 & 1 & 0.77 \\ 0.77 & 0.76 & 0.77 & 1 \end{bmatrix},$$

with rows and columns representing fatigue, physical function, ability to participate in social roles and activities, and the COPD-SIB, respectively. The item parameters are presented in Supplement 2. The discrimination parameters were relatively high for all domains (range: 0.82–5.40), which is quite common for clinical measures [32], and the β_{ij} parameters showed a good spread (range: −7.57 to 7.67). Measurement precision for θ-estimates was excellent (RMSE < 0.3 for all domains). The direction of bias was in line with the expected shrinkage (which is the result of the implementation of a Bayesian estimator): positive θ-values tended to be slightly underestimated and low negative θ-values tended to be overestimated. See Supplement 3 for RMSE and bias plots.

Data generation and CAT simulations

CAT simulations were run with the package ShadowCAT [33] in R [34]. To evaluate the item usage rates of all individual items in our item bank, responses were generated based on 21000 vectors of pre-specified θ-values—1000 for every increment of 0.2 on the multidimensional θ-scale between values −2 and 2. The Maximum A Posteriori (MAP) estimator was used in all simulations to estimate θ, at all stages of the CAT. The covariance matrix Φ estimated using the multidimensional GRM was used as a prior. Following Segall [19], item selection was based on the value of the determinant of the posterior information matrix. Diao and Reckase [35] refer to this item selection method as *Bayesian Volume Decrease*, whereas Yao [36] simply abbreviates it as *Volume* or *Vm*. One random item per domain was administered at the start in order to obtain initial θ-values to initialize the CAT. The CAT was terminated, when the termination rule (threshold standard error of measurement $SE(\theta) < 0.316$)[3] was met for all four domains. Item selection for a particular dimension was terminated, when the *SE*-threshold had been met for that dimension.

Outcome variables

The outcome variables in this study were overuse and active domain/bank size, all conditional on θ. Each of the outcome variables will be reported by domain as well as across domains (i.e., at item bank level). An item was considered overused when its usage rate was higher than the expected item usage rate,[4] defined as the average test length for a given θ-value divided by the total bank size (194). Active domain/bank size was calculated as total domain or bank size minus items that were never used in the respective domain or overall bank.

[3] In unidimensional models, an *SE*-value of 0.316 corresponds to a local reliability of 0.90 when a variance of 1 for θ is assumed (see, e.g., [37]).

[4] Note that *usage* was operationalized as selection in the adaptive part of the CAT (items selected at random to be used as start items to initialize the CAT were ignored in calculating the outcome variables).

Results

The results of the CAT simulations are summarized in Tables 1 and 2, Fig. 1, and Supplement 4. Table 1 illustrates that there was—as could be expected—quite some diversity in active bank size across the different θ-values. For average θ-values, the overall active bank size was 9–10%; this number quickly increased as θ-values became more extreme. For values of -2 and $+2$, the overall active bank size increased fourfold to 39–40%! Unsurprisingly, CATs for more extreme θ-values (-2 and $+2$) were generally longer than for less extreme values (average length of 20.5 and 18.9 versus 14.1, 13.3, and 13.0 for θ-values; -1, 0, and $+1$, respectively). However, the active bank size increased at a steeper rate than the test length, for increasing absolute θ-values. There was also considerable diversity in active bank size across domains. For average θ-values, the active domain size for fatigue and physical function was 5–6%, compared to 9–11% for ability to participate in social roles and activities, and 17% for the COPD-SIB. For extreme θ-values, almost all ability to participate in social roles and activities items were used; this finding can be directly linked to the item parameter distributions for this bank (high discrimination parameters combined with broad coverage on the θ-scale); see Fig. 1.

Comparing Table 2 (percentage of overused items) to Table 1 (active bank size) shows that—for the total bank and average θ-values—overused items dominated the active part of the multidimensional item bank; there was 78% overlap between overused items and active bank size. For more extreme θ-values, the overused items made up a much smaller part of the active bank size: here the overlap was only 35%.

Figures 1–4 in Supplement 4 illustrate that there are 12 items that have relatively high item usage rates over a wide range of θ-values: FATIMP1 ("To what degree did you have to push yourself to get things done because of your fatigue?"), FAMTIMP9 ("How often did your fatigue make it difficult to plan activities ahead of time?"), FATIMP29 ("How often were you too tired to leave the house?"), PFB1 ("Are you able to climb up five steps?"), PFB44 ("Does your health now limit you in doing moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf?"), SRPPER20 ("I have trouble doing all of the activities with friends that are really important to me"), SRPPER23 ("I have trouble doing all of my usual work (include work at home)"), SGRQ12 ("Please, indicate whether the following activity causes shortness of breath. If the weather influences your complaints, assume the weather conditions are favorable, when

Table 1 Active bank size (expressed in %) for θ-values ranging between -2 and $+2$

θ	Total bank	Fatigue	Physical function	Social roles	COPD-SIB
−2	40	26	33	71	39
−1.8	24	20	16	37	30
−1.6	24	20	17	34	28
−1.4	19	14	16	23	26
−1.2	17	10	13	23	26
−1	15	10	14	20	20
−0.8	13	8	13	14	20
−0.6	10	8	8	11	15
−0.4	9	6	6	11	15
−0.2	9	6	6	11	15
0	9	6	5	9	17
0.2	9	6	5	11	17
0.4	9	6	5	9	17
0.6	10	8	5	11	17
0.8	9	10	3	9	17
1	10	12	3	11	17
1.2	15	18	6	20	22
1.4	17	20	6	23	24
1.6	19	22	10	20	26
1.8	24	26	14	31	30
2	39	30	16	97	37
Full bank size[a]	194	50	63	35	46

[a] Number of available calibrated items in each domain/the total bank

Table 2 Overused items (expressed in %) for θ-values ranging between -2 and $+2$

θ	Total bank	Fatigue	Physical function	Social roles	COPD-SIB
−2	15	10	13	17	22
−1.8	14	8	13	17	20
−1.6	12	6	11	17	17
−1.4	11	6	8	17	17
−1.2	9	6	8	9	15
−1	8	6	6	6	15
−0.8	8	6	6	6	13
−0.6	7	6	5	6	13
−0.4	7	6	5	6	11
−0.2	7	6	5	6	11
0	7	6	5	6	11
0.2	7	6	3	9	13
0.4	8	6	3	9	15
0.6	7	6	3	6	15
0.8	7	6	3	6	15
1	8	8	3	6	17
1.2	9	10	3	6	17
1.4	10	12	3	9	17
1.6	11	12	6	11	17
1.8	13	12	8	20	15
2	14	14	8	17	20

Overused items are defined as items whose usage rate exceeded the expected usage rate (average test length for a given θ-value divided by the total bank size)

you answer this question. Getting washed or dressed"), SGRQ13 ("Please, indicate whether the following activity causes shortness of breath. If the weather influences your complaints, assume the weather conditions are favorable, when you answer this question. Walking around the home."), SGRQ26 ("I get afraid or panic when I cannot get my breath."), SGRQ42R1a ("My breathing problems make it difficult to do light gardening, such as weeding."), and SGRQ42R1b ("My breathing problems make it difficult to do things such as dancing, playing golf, or playing bowls.").

Some items, such as CSIB13 ("It frustrated me that I couldn't do everything I wanted to do anymore") and SGRQ31 ("Everything seems too much of an effort."), show two peaks; something typical for polytomous data. Polytomous items have more than one β parameter and thus cover a wider θ-range. A polytomous item can have more than one peak in its item information function, which would translate into more than one peak in the item usage plot. Longer CATs are needed to obtain reliable estimates of very low or high θ-values, which explains why as many as 38 items show relatively high item usage rates for low or high θ-values only.

In Fig. 1, the item step parameters are plotted against the discrimination parameters for each domain. The figure clearly shows that within each domain, the items with the highest discrimination values had the highest item usage rates. These items typically covered a wide range of θ-values.

Discussion

In this study, we evaluated active bank size and item overuse/overexposure in a recently developed MCAT designed to measure HRQL in COPD patients using four correlated domains. Three generic PROMIS domains were used: the PROMIS domains fatigue, physical function, and ability to participate in social roles and activities [25, 26]; as well as a COPD-specific item bank (the COPD-SIB) which was recently developed [16]. We found that, for average latent trait values, the overall active bank size was 9–10%; compared to 39–40% for more extreme latent trait values (-2 and $+2$). Furthermore, as expected, domains with highly discriminating items were overrepresented in the active part of the multidimensional bank. For average latent trait values, the active part of the bank was almost entirely populated by overused items. In contrast, for more extreme latent trait values, the active part of the multidimensional bank was dominated by underused items. The number of items that showed good item usage and covered

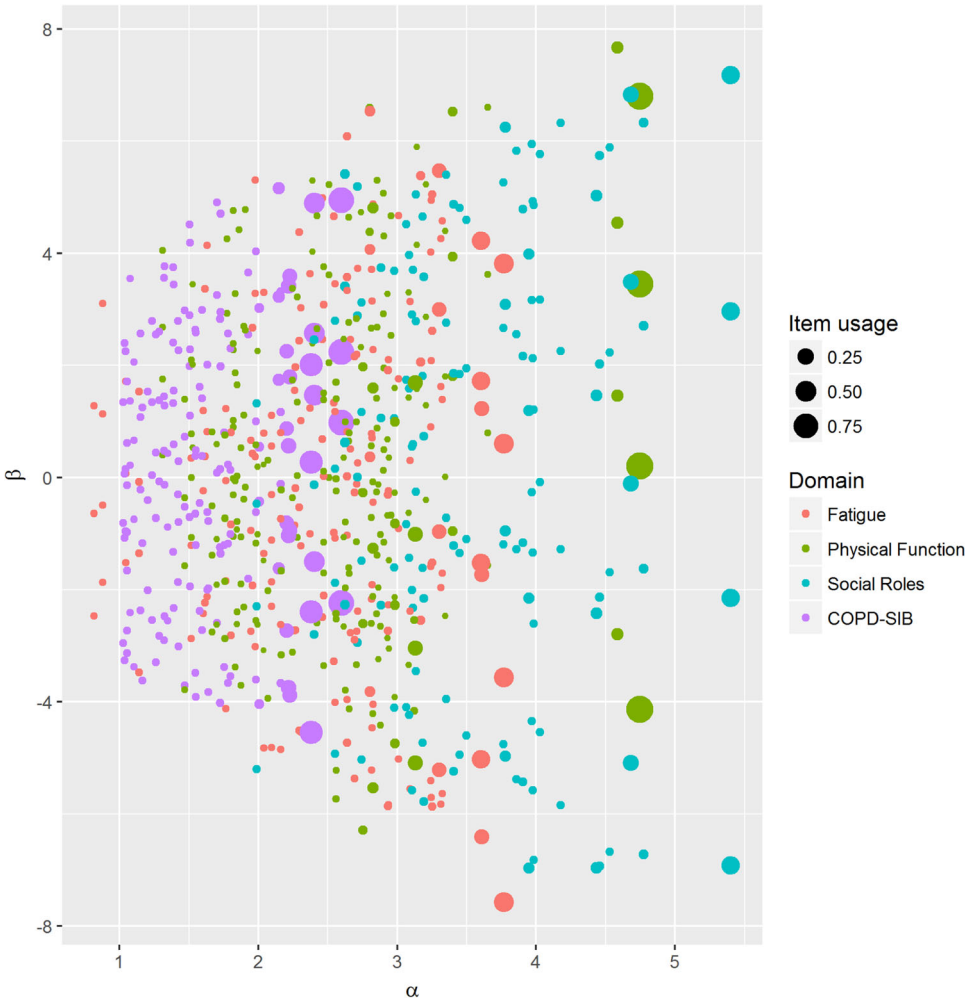

Fig. 1 Scatterplot with discrimination values on the *x*-axis and *β*-parameter values (related to difficulty) on the *y*-axis. *Dots* represent item steps. The size of the *dots* increases as a function of item usage rate. See online for color version

almost the entire latent trait range varied between 2 (physical function) and 5 (COPD-SIB) per domain.

We used a multidimensional item bank consisting of 194 items (35–63 items per domain). Given that we developed a MCAT without content constraints and with no exposure control, our results indicate that the MCAT was working as intended: for average latent trait values, a small number of highly discriminating items was selected; for more extreme values, the item bank usage was more balanced. However, our results also showed that a relatively large part of the multidimensional item bank was never used (60%). The active part of the bank consisted of 77 items at most, across the four domains. This may indicate that—if these findings can be generalized—roughly 19 polytomous items per domain might suffice, when developing a multidimensional bank populated by items with high discrimination parameters that adequately cover the latent trait range of interest, and with high correlations among domains. Research focusing on unidimensional CATs has shown that CATs based on polytomous rather than dichotomous items can be performed with substantially smaller item banks; an item bank of 30 items may be sufficient for polytomously scored health outcomes [38, 39]. Our results suggest that MCAT potentially requires smaller item banks than UCAT. It would be interesting to study this further in a future study.

Item usage has received little attention in the field of clinical (psychological/health) measurement so far. One exception concerns developing IRT/CAT-based short forms. Several authors have suggested that CAT simulations can be used to select the most appropriate items for inclusion in a short form [40–42]. In these studies, typically the entire item pool is administered, after which the rank order in which the items were administered is calculated and averaged over all simulees. The "best" items (items with the lowest average CAT presentation ranks) would then be selected for the short form [41]. Items for the newest PROMIS short forms were selected based on the maximum interval information and CAT simulations (highest average administration rank) [43], making their measures easily accessible in situations where CAT may not be feasible. Because static short forms will be typically targeted at a relatively wide latent trait range, they are

relatively long compared to CATs, especially for respondents with average latent trait values. Furthermore, although a short form may achieve adequate measurement precision for average to moderately high latent trait scores, CATs provide much better precision at the extremes [41, 42]. Our results showed how active bank size and the rate of overused items also depended on latent trait values. In other words, which items are the "best" items (in terms of administration rank/usage) depends largely on the respondent's latent trait values. This is not something that can be satisfactorily addressed in a short form.

Another topic which has received little attention in our field is the influence of capitalization on item calibration error. Since the item selection criterion most frequently used is a direct function of the discrimination parameter, item selection is sensitive to large standard errors of discrimination parameters [44, 45]. Typically, extreme discrimination parameter estimates tend to be associated with larger standard errors [46]. Furthermore, the smaller the selection ratio (CAT length divided by total item bank), the larger the danger of capitalization on chance [47]. Capitalization on item calibration error may lead to overestimation of test information and underestimation of the standard errors of latent trait estimates [46]. In this light, having a small set of items with very high item usage rates (and a large set not being used at all) may be worrying, regardless of the issue of test security. In this study, we did find a strong correlation (0.82) between estimated discrimination parameters and their respective standard errors. However, penalizing items with the highest discrimination parameter estimates (for example, by increasing the estimates by 1 or 2 times their corresponding standard error), would have had a very insubstantial effect on their ranking (data not shown). This being said, if we would have penalized items with relatively high standard errors during the CATs, test length would most likely have been somewhat longer, and subsequently the active size of the item bank would also have been larger. Since estimates are typically (also in our case; data not shown) more precise when using a multidimensional rather than unidimensional IRT models to calibrate the items, the impact of item calibration error can be expected to be smaller than if we had used separate unidimensional CATs. Research investigating the potential protective effect of multidimensional IRT and CAT on the consequences of capitalization on item calibration error is needed.

Conclusion

With this study, we extended the literature on item usage rates to multidimensional health measurement. We showed what happens when realistic CAT settings (typical for health measurement) are used: a relatively small number of highly discriminating items is selected. Currently, PROMIS item banks differ widely in length. Our results strengthen the claim that relatively short item banks may suffice when using polytomous items (and no content constraints/exposure control mechanisms), especially when using MCAT. This may be particularly relevant to item bank developers. However, if researchers or clinicians want to be able to influence the content (to ensure validity), different item selection procedures are necessary; in such instances, a larger item bank will be needed.

Acknowledgements 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. We thank all patients who participated in the study. This study was supported by Grant #3.4.11.004 from Lung Foundation Netherlands.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval The ethical review board of the University of Twente approved the study. This study did not need approval of the Medical Ethical Review Board, according to European regulations. All procedures performed in studies involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent All patients gave informed consent.

References

1. Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic Publishers.
2. Embretson, S. E., & Reise, S. (2000). *Item response Theory for psychologists*. Mahwah, NJ: Erlbaum.
3. Devine, J., Otto, C., Rose, M., Barthel, D., Fischer, F., Muhlan, H., et al. (2015). A new computerized adaptive test advancing the measurement of health-related quality of life (HRQoL) in children: The Kids-CAT. *Quality of Life Research, 24*(4), 871–884. doi:10.1007/s11136-014-0812-7.
4. Sunderland, M., Slade, T., Krueger, R. F., Markon, K. E., Patrick, C. J., & Kramer, M. D. (2016). Efficiently measuring dimensions

of the externalizing spectrum model: Development of the externalizing spectrum inventory-computerized adaptive test (ESI-CAT). *Psychological Assessment*. doi:10.1037/pas0000384.
5. Devine, J., Otto, C., Rose, M., Barthel, D., Fischer, F., Mulhan, H., et al. (2015). A new computerized adaptive test advancing the measurement of health-related quality of life (HRQoL) in children: The Kids-CAT. *Quality of Life Research, 24*(4), 871–884. doi:10.1007/s11136-014-0812-7.
6. Stucky, B. D., Huang, W., & Edelen, M. O. (2015). The psychometric performance of the PROMIS smoking assessment toolkit: Comparisons of real-data computer adaptive tests, short forms, and mode of administration. *Nicotine & Tobacco Research*. doi:10.1093/ntr/ntv083.
7. Rose, M., Bjorner, J. B., Fischer, F., Anatchkova, M., Gandek, B., Klapp, B. F., et al. (2012). Computerized adaptive testing–ready for ambulatory monitoring? *Psychosomatic Medicine, 74*(4), 338–348. doi:10.1097/PSY.0b013e3182547392.
8. 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*(9), 1223–1229. doi:10.1016/j.apmr.2006.05.018.
9. Nikolaus, S., Bode, C., Taal, E., Oostveen, J. C., Glas, C. A., & 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*(10), 1175–1183. doi:10.1016/j.jclinepi.2013.05.010.
10. Makransky, G., Mortensen, E. L., & Glas, C. A. (2013). Improving personality facet scores with multidimensional computer adaptive testing: an illustration with the NEO PI-R. *Assessment, 20*(1), 3–13. doi:10.1177/1073191112437756.
11. Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331–354. doi:10.1007/bf02294343.
12. Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*(5), 295–316. doi:10.1177/0146621604265938.
13. Paap, M. C. S., Kroeze, K. A., Glas, C. A. W., Terwee, C. B., van der Palen, J., & Veldkamp, B. P. (Under review). Measuring patient-reported outcomes adaptively: Multidimensionality matters!
14. Paap, M. C. S., Bode, C., Lenferink, L. I. M., Groen, L. C., Terwee, C. B., Ahmed, S., et al. (2014). Identifying key domains of health-related quality of life for patients with chronic obstructive pulmonary disease: The patient perspective. *Health Qual Life Outcomes, 12*(1), 106. doi:10.1186/s12955-014-0106-3.
15. Paap, M. C. S., Bode, C., Lenferink, L. I. M., Terwee, C. B., & van der Palen, J. (2015). Identifying key domains of health-related quality of life for patients with chronic obstructive pulmonary disease: Interviews with healthcare professionals. *Quality of Life Research, 24*(6), 1351–1367. doi:10.1007/s11136-014-0860-z.
16. 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*(97), 1–15. doi:10.1186/s12955-016-0500-0.
17. Veldkamp, B. P., & van der Linden, J. W. (2000). Designing item pools for computerized adaptive testing. In J. W. Van der Linden & A. W. C. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic Publishers.
18. Hau, K.-T., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*(3), 249–266.
19. 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: Kluwer Academic Publishers.
20. Barrada, J. R., Veldkamp, B. P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*(1), 58–73. doi:10.1177/0146621608315329.
21. Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, *5*(8). Retrieved 8 Nov 2016 from http://www.jtla.org.
22. Fayers, P. M., & Hays, R. (2005). *Assessing quality of life in clinical trials: Methods and practice* (2nd ed.). Oxford: Oxford University Press.
23. Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23. doi:10.1177/0146621697211001.
24. Paap, M. C. S., Bode, C., Lenferink, L. I. M., Terwee, C. B., & Palen, J. (2015). Identifying key domains of health-related quality of life for patients with chronic obstructive pulmonary disease: Interviews with healthcare professionals. *Quality of Life Research, 24*(6), 1351–1367. doi:10.1007/s11136-014-0860-z.
25. Cella, D. F., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (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*(11), 1179–1194. doi:10.1016/j.jclinepi.2010.04.011.
26. Terwee, C. B., Roorda, L. D., de Vet, H. C., Dekker, J., Westhovens, R., van Leeuwen, J., et al. (2014). Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). *Quality of Life Research, 23*(6), 1733–1741. doi:10.1007/s11136-013-0611-6.
27. Ninot, G., Soyez, F., & Prefaut, C. (2013). A short questionnaire for the assessment of quality of life in patients with chronic obstructive pulmonary disease: Psychometric properties of VQ11. *Health Qual Life Outcomes, 11,* 179. doi:10.1186/1477-7525-11-179.
28. Meguro, M., Barley, E. A., Spencer, S., & Jones, P. W. (2007). Development and validation of an improved, COPD-specific version of the St. *George Respiratory Questionnaire. Chest, 132*(2), 456–463. doi:10.1378/chest.06-0702.
29. Vidotto, G., Carone, M., Jones, P. W., Salini, S., & Bertolotti, G. (2007). Maugeri Respiratory Failure questionnaire reduced form: A method for improving the questionnaire using the Rasch model. *Disability and Rehabilitation, 29*(13), 991–998. doi:10.1080/09638280600926678.
30. Maille, A. R., Koning, C. J., Zwinderman, A. H., Willems, L. N., Dijkman, J. H., & Kaptein, A. A. (1997). The development of the 'Quality-of-life for Respiratory Illness Questionnaire (QOL-RIQ)': A disease-specific quality-of-life questionnaire for patients with mild to moderate chronic non-specific lung disease. *Respiratory Medicine, 91*(5), 297–309.
31. Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO 2.1 [computer software]*. Lincolnwood: Scientific Software.
32. Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*(1), 27–48. doi:10.1146/annurev.clinpsy.032408.153553.
33. Kroeze, K. A., & de Vries, R. (2015). ShadowCAT: Multidimensional Computer Adaptive Testing with the Shadow Testing routine. Retrieved 3 April 2017 from https://github.com/Karel-Kroeze/ShadowCAT/.

34. R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
35. 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 8 Nov 2016 from https://www.psych.umn.edu/psylabs/CATCentral/.
36. Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*(1), 3–23. doi:10.1177/0146621612455687.
37. Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*(1), 1–18.
38. Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5–22. doi:10.1177/014662169501900103.
39. Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229–255). New York, NY: Routledge.
40. Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*(4), 347–364. doi:10.1177/107319110000700404.
41. Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*(1), 125–136. doi:10.1007/s11136-009-9560-5.
42. Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2007). Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms. *Applied Psychological Measurement, 31*(5), 412–429. doi:10.1177/0146621606297314.
43. PROMIS. Retrieved November 8, 2016 from https://www.assessmentcenter.net/Manuals.aspx.
44. van der Linden, W. J., & Glas, C. A. W. (2001). Cross-validating item parameter estimation in adaptive testing. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 205–219). New York, NY: Springer.
45. Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement, 37*(2), 123–139. doi:10.1177/0146621612469825.
46. Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*(2), 143–155.
47. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.