Qual Life Res (2017) 26:2949–2959
DOI 10.1007/s11136-017-1627-0

Measuring everyday functional competence using the Rasch assessment of everyday activity limitations (REAL) item bank

Martijn A. H. Oude Voshaar[1] · Peter M. ten Klooster[1] · Harald E. Vonkeman[1,2] · Mart A. F. J. van de Laar[1,2]

Accepted: 15 June 2017 / Published online: 21 June 2017

Abstract

Objective Traditional patient-reported physical function instruments often poorly differentiate patients with mild-to-moderate disability. We describe the development and psychometric evaluation of a generic item bank for measuring everyday activity limitations in outpatient populations.

Study design and setting Seventy-two items generated from patient interviews and mapped to the International Classification of Functioning, Disability and Health (ICF) domestic life chapter were administered to 1128 adults representative of the Dutch population. The partial credit model was fitted to the item responses and evaluated with respect to its assumptions, model fit, and differential item functioning (DIF). Measurement performance of a computerized adaptive testing (CAT) algorithm was compared with the SF-36 physical functioning scale (PF-10).

Results A final bank of 41 items was developed. All items demonstrated acceptable fit to the partial credit model and measurement invariance across age, sex, and educational level. Five- and ten-item CAT simulations were shown to have high measurement precision, which exceeded that of SF-36 physical functioning scale across the physical function continuum. Floor effects were absent for a 10-item empirical CAT simulation, and ceiling effects were low (13.5%) compared with SF-36 physical functioning (38.1%). CAT also discriminated better than SF-36 physical functioning between age groups, number of chronic conditions, and respondents with or without rheumatic conditions.

Conclusion The Rasch assessment of everyday activity limitations (REAL) item bank will hopefully prove a useful instrument for assessing everyday activity limitations. T-scores obtained using derived measures can be used to benchmark physical function outcomes against the general Dutch adult population.

Keywords Item response theory · Item bank · Computerized adaptive testing · Physical function · Activity limitations · Rasch

Introduction

Physical function can be defined as a person's ability to engage in activities that require physical movement and exertion with the purpose of performing everyday tasks or needs. It is a central component of health-related quality of life and a key outcome in its own right across many medical conditions [1–7]. The level of self-reported physical disability experienced by individuals has been shown to vary widely within conditions for which physical function is commonly assessed [8, 9]. A fixed-length questionnaire that includes items relevant to each of these levels would have to include many questions that would not all be relevant for individual respondents. Since short questionnaires are usually preferred for feasibility reasons, it has proven challenging to develop fixed-length questionnaires

Electronic supplementary material The online version of this article (doi:10.1007/s11136-017-1627-0) contains supplementary material, which is available to authorized users.

✉ Martijn A. H. Oude Voshaar
a.h.oudevoshaar@utwente.nl

[1] Arthritis Center Twente, Department of Psychology, Health and Technology, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands

[2] Arthritis Center Twente, Department of Rheumatology and Clinical Immunology, Medical Spectrum Twente, Enschede, The Netherlands

that adequately measure the variety of physical function levels that might occur within a population of interest [10–12].

Item response theory (IRT)-based item banking has been promoted as a powerful solution for overcoming well-known limitations of fixed-length instruments, such as floor and ceiling effects or insensitivity to change [13–15]. IRT is a framework for modeling item response data in which items and respondents are located on a common scale [16, 17]. Once the scale location of a set of items has been estimated precisely, any number and combination of items in the resulting item bank can be used to estimate the scale location of future respondents from the same population. Computerized adaptive testing (CAT) is an application of IRT that utilizes this feature to optimize measurement precision of individual assessments and reduce the number of items that need to be administered for a precise estimate, in which an algorithm selects items that best match the individual's estimated level of the measured trait in real-time [18]. This could yield shorter, potentially equiprecise assessments, depending on the characteristics of the item bank.

However, studies in outpatient settings have shown currently available physical function item banks to poorly differentiate patients with mild-to-moderate disability [19–21] and most items to target lower levels of function [22–26]. This likely reflects that much of their content is derived from previously validated questionnaires [23, 27]. Many such questionnaires were developed for use in populations with severe disabilities and focus on basic activities of daily living (ADL) that are essential for independent self-care such as toileting, grooming, and eating, as well as basic mobility functions such as getting in and out of bed [28–30].

Since the precision with which scores can be estimated is optimal if an instrument's items are well matched with trait levels of the respondents, a balanced coverage of the full spectrum of physical function would likely further improve measurement performance of physical function item bank applications in a variety of settings. Previous studies suggest that items reflecting various activities that people routinely engage in as part of managing their domestic responsibilities better match the levels of disability typically experienced by outpatients compared with the basic ADL. [31–37]. However, item content analyses have shown that such activities are infrequently included in the validated instruments that are now in widespread use [38–43]. The objective of the present work was to develop a new item bank that assesses disability in complex activities typically encountered in the daily lives of independently living individuals and to provide an initial evaluation of its measurement properties using an empirical CAT simulation.

Methods

Item pool development

The items were designed to capture the health concepts included in the International Classification of Functioning, Disability and Health (ICF), domestic life domain, using content that was derived as much as possible from people with first-hand experience with physical disability. An initial list of activities was derived from the responses obtained in a survey among a convenience sample of 103 consecutive patients attending a rheumatology outpatient clinic in the Netherlands. Characteristics of these patients are summarized in supplementary Table 1. Patients were asked: "Many people with [medical condition] experience difficulties in performing daily physical activities, such as climbing stairs or gardening. When you think about your [medical condition], what limitations in daily activities bother or upset you the most?" We linked the activities mentioned by patients to the ICF [44] and selected the 143 (39%) activities that were mapped to any third-level code within the ICF domestic life chapter. The process of binning and winnowing, which involves grouping together related concepts and deleting redundant content from these groups [27], was applied to these 143 activities, yielding a list of 57 sufficiently distinct activities. The degree to which this list comprehensively covered all ICF domestic life second-level codes was reviewed and 15 additional items were written by the project team to ensure a balanced coverage of the ICF domestic life second-level codes. We used slightly adapted versions of the PROMIS physical function item bank and health assessment questionnaire disability index (HAQ) item stems ("are you physically able to [activity]?") and response options (without any difficulty/with a little difficulty/with some difficulty/with a lot of difficulty/cannot do), since these have been extensively tested in qualitative research and are widely used [27, 30].

Calibration study participants and data collection procedures

The item pool was administered to 1128 adults enlisted in the Longitudinal Internet Studies for the Social sciences (LISS) panel, an academic research panel administered by Tilburg University in collaboration with Statistics Netherlands, for which a true probability sample was drawn from the Dutch population registers [45]. The LISS panel has been shown to approximate national statistics on the total non-incarcerated population in the Netherlands better than 18 online panels based on non-probability, self-selected samples in terms of coverage of sociodemographic

Table 1 Sample characteristics

Age, years mean (SD) min–max	50.36 (17.99) 16–97
Sex, *n* (%)	
Male	524 (46.5%)
Female	604 (53.5%)
Educational level, *n* (%)[a]	
Low	314 (28%)
Middle	416 (37%)
High	394 (35%)
Occupational status, *n* (%)	
Remuneratively employed	546 (48.4%)
Pensioned	261 (23.1%)
Student	107 (9.5%)
Housekeeper	85 (7.5%)
Looking for work	54 (4.7%)
Unable to work	41 (3.6%)
Voluntarily employed	29 (2.6%)
Other	5 (0.4%)
Self-reported diagnosis of chronic condition, *n* (%)	
Osteoarthritis	162 (14.4%)
Diabetes mellitus	65 (5.8%)
Asthma	65 (5.8%)
COPD	40 (3.5%)
Depression	62 (5.5%)
Rheumatoid arthritis	31 (3.0%
Fibromyalgia	(20 (1.8%)
Stroke	6 (0.5%)
Hypertension	211 (18.7%)
Migraine headaches	59 (5.2%)
Any rheumatic condition	190 (17%)
Any chronic condition	491 (43.5%)
T-score[a], mean (SD) min–max	50.62 (17.54) 24.62–89.58
SF-36 Physical functioning, Mean (SD) min–max	83.06 (25.28) 0–100

COPD chronic obstructive pulmonary disease; *SF-36* 36-Item Short-form Health Survey; T-score is a item bank-derived physical function estimate, transformed to a scale where mean = 50 (SD = 10), with higher values indicating better function)

[a] According to UNESCO International standard classification of education

characteristics [46]. Representativeness of the panel has also been demonstrated to approximate a major national face-to-face survey (the Dutch parliamentary election study) conducted by Statistics Netherlands on all tested variables except with respect to coverage of the elderly (>70 years old) and the non-internet population [47]. The module that respondents completed for our study contained the item pool of 72 items, basic demographic information, and the SF-36 physical functioning scale (PF-10), which contains 10 items measuring perceived current limitations in a variety of physical activities on a 3-point response scale. Scores are summed and linearly transformed to range between 0 and 100, with higher scores indicating better functioning. Previous studies have shown the PF-10 to have favorable measurement properties in a variety of settings [48].

Respondents were also asked if they had ever been diagnosed by a physician with any of the following chronic conditions that are prevalent in The Netherlands according to the Dutch National Institute for Public Health and the Environment (www.volksgezondheidenzorg.info): osteoarthritis, diabetes, asthma, chronic obstructive pulmonary disease, depression, rheumatoid arthritis, fibromyalgia, hypertension, stroke, congenital heart disease, or migraine. Finally, respondents were asked to rate the level of difficulty they had filling out the questionnaire and the clarity of the survey in

general, including the PF-10 and the other items using 5-point Likert scales with response options ranging from not at all difficult/unclear to extremely difficult/clear.

Assessment of item quality and IRT assumptions

IRT models for ordered polytomous data assume that the score on a questionnaire item is explained by a (set of) latent trait(s) and the relationship between the observed score and the trait(s) can be described by a monotonically increasing function [16]. The dimensionality of the item pool and the assumption that the expected scores of individual items are continuously non-decreasing across the physical function continuum were examined preceding the IRT analysis. Since all items were newly developed for this study, we used exploratory factor analysis (EFA) with robust weighted least squares estimation on polychoric correlations in Mplus 7.11 to explore the latent structure of the item pool. The number of factors to retain was decided after examining the eigenvalues (i.e., the variances of the extracted factors), and model-data fit was assessed using proposed cut-off values for the fit indices provided by Mplus [49]. Based on the results of this analysis, we decided if a unidimensional or multidimensional IRT model would be used. We excluded items from EFA for which individual response options were selected by <20 respondents since it would be difficult to obtain good estimates of the IRT intersection parameters (described below) at a later stage. Items for which >80% of responses were in either extreme end of the 5-point response scale were also excluded.

The assumption that expected item scores are monotonically increasing throughout the range of trait values was examined using the check for monotonicity function in the Mokken R-package [50]. In this procedure, respondents are grouped based on their rest score (i.e., total score on all items minus the item under consideration) and it is examined if the percentages of respondents endorsing each of an item's response options are continuously non-decreasing for all levels of the restscore. For each item, the number of violations of monotonicity was counted, as well as the number of statistically significant violations. In case of significant violations, the 'crit' statistic quantifies the magnitude of violation. Values >40 are considered to be problematic [51].

IRT modeling

We considered a series of generalizations of the partial credit model (PCM), which is a Rasch-type model for ordered polytomous data in which the item and person characteristics are located on the same logit scale. The PCM specifies that for an item i with $k = 1..\ m$ score categories, the log odds of a score in category k instead of $k-1$ by a person n depend only on the distance on the measurement continuum between that respondent's physical function level θ_n and a parameter β_{ik} that reflects the location on the scale where a response in both categories is equally likely:

$$\mathrm{LN}\frac{P_{nik}}{P_{ni(k-1)}} = \theta_n - \beta_{ik}.$$

The PCM can be generalized by introducing a discrimination parameter that allows the amount of change in the log odds for one unit of change in θ_n to be different between items [52]. If needed, this model can be further generalized to accommodate multiple dimensions. Since these models are nested, we compared their relative fit using a likelihood ratio test and we used an item-level Lagrange multiplier (LM) test [53].

After a model was chosen, we examined the presence of age, educational level, and sex-related differential item functioning (DIF) using LM statistics and associated effect size (ES) statistics, calculated as the expected value of the residuals across three score level groups for item fit and across sociodemographic subgroups considered in the DIF analyses. The residuals were divided by the maximum attainable item score, such that an ES of, for example, 0.10 indicates that the observed average score was 10% different from its expectation under the model [54]. For age, the sample was split into three groups (0–39, $n = 354$; 40–59, $n = 364$; 60+, $n = 410$) and for educational level respondents were classified in three groups in accordance with the International Standard Classification of Education [55]. Pronounced DIF was defined as a p value for the LM test <0.01 in combination with an ES of >5%. The IRT models we considered all assume that the item responses depend only on the value(s) of the latent trait(s) in the model. Local stochastic independence was evaluated using Yen's $Q3$ statistic (i.e., the correlation between fit residuals) [56]. Items with an absolute $Q3 > \pm\ 0.20$ were flagged for possible local stochastic dependence [17, 56, 57]. The impact of local dependence was examined by examining the change in parameter estimates after omitting locally dependent items. All IRT analyses were performed using the MIRT package [58].

Analysis of measurement properties

A preliminary evaluation of the performance of item bank-derived measures was performed by using the responses to the calibrated items as input for an empirical CAT simulation. The first item to be administered was selected to maximize information at the mean of the distribution of trait levels. The maximum posterior weighted information criterion [59] was used for subsequent item selection, using

the expected a posteriori procedure for interim latent score estimation. A standard normal prior was used for both interim item selection and interim score estimation. Final estimates and their standard errors were obtained using the maximum likelihood (ML) method.

Measurement precision

The measurement precision of the CAT algorithm with various numbers of administered items was compared with that of the PF-10 and administration of 10 random items from the item bank by plotting the standard errors of the ML estimates of the different measures across the various CAT score levels. We also compared floor and ceiling effects for all administered items, defined as the percentage of persons that selected the highest and lowest response option, respectively. We examined floor and ceiling effects in the total study population as well as in the subpopulation with ≥1 self-reported chronic condition and in the subpopulation with any self-reported rheumatic condition (i.e., rheumatoid arthritis, osteoarthritis, or fibromyalgia).

Sensitivity

We compared the ability of the instruments to differentiate between respondents with self-reported osteoarthritis, rheumatoid arthritis, or fibromyalgia (i.e., with a rheumatic condition) versus those without any rheumatic condition. We also compared scores between respondents who self-reported to have been diagnosed with 0, 1, or >1 chronic conditions and respondents in different age groups. We hypothesized that all measures would discriminate between these groups and that T-scores and PF-10 scores would decrease with the number of chronic conditions and age. To compare discriminative ability of the different instruments, relative efficiency coefficients were obtained [60].

Results

Sample characteristics

Some characteristics of the sample are summarized in Table 1. Physical function scores were high on average, but respondents of all functional levels assessed by the PF-10 were represented in the sample.

IRT assumptions

Twenty-five items with ≥80% of responses in the "with no difficulties" category were excluded. All remaining items had >20 responses in each response category. EFA on the remaining 47 items revealed three factors with eigenvalues >1 and a strong first factor, with a first-to-second eigenvalue ratio of 13.14. The first three eigenvalues were 34.95, 2.66, and 1.38 and the first factor explained 77% of the total variance. Furthermore, all items had significant ($p < 0.05$) loadings on the first factor, generally in excess of 0.77, except for item 2 (sewing clothes by hand using needle and thread), for which the loading on the first factor was 0.53. Except for RMSEA (0.09), all fit indices suggested a good fit (TLI = 0.96, CFI = 0.96), SRMSR = 0.10 and did not improve much for the two-factor solution, further supporting a unidimensional measurement model. For the one-factor model, only 1% of item pairs had residual correlations >0.20. Based on these results, we concluded that a unidimensional item response model was probably most suitable for these data. In the Mokken analysis, six items were identified with each one violation of monotonicity. However, none of these violations were statistically significant and the 'crit' statistic was <40 in all cases (Max *crit* = 7). All items had scalability coefficients >0.30, indicating that respondents can be ordered on the latent continuum using the total score (Table 2).

IRT model fitting

In an initial evaluation of all 46 items, the LR test yielded a value of $\chi^2 = 646{,}826$, df = 46, $p < 0.01$, indicating superior fit for the GPCM compared with PCM. However, for both models, ESs were generally of small magnitude (i.e., ES <2%) and the root mean squared difference ES was <0.6%, suggesting that both models performed similarly in terms of reproducing the item response data. Only item 2 met the criteria for substantial lack of fit in the PCM calibration (LM = 107.39 $p < 0.01$, ES = 5.8%), but not the GPCM calibration. This was likely due to its weak relation with the general factor observed earlier in the EFA.

Although GPCM performed better according to the LR test, we elected to delete item 2 and proceed with the more parsimonious PCM. Inspection of the matrix with $Q3$ statistics revealed that 6% of the item pairs had an absolute $Q3$ >0.20. After removing the item with the highest average $Q3$, item threshold parameters of the remaining items changed by a maximum of 0.13 and the mean absolute difference between threshold parameters in the original calibration and the re-estimated item parameters was 0.04 (SD = 0.03). After 10 items were removed, the mean absolute difference was 0.04 (SD = 0.01), suggesting a small impact of local dependence on the parameter estimates. For age or educational attainment, none of the items had substantial DIF in the PCM calibration. However, for 2 items, involving preparing meals and washing clothes by hand, scores were >5% lower for male respondents than

Table 2 Item characteristics of the final 41 items of the Rasch Everyday Activities Limitations item bank

ICF code	Item	β_1	β_2	β_3	β_4	Effect size (%)	H_{ij} coefficient
d65506	Cleaning the oven	29	13	9	0	0.2	0.68
d6201	Clothes shopping	27	21	10	0	0.4	0.69
d65502	Turning in light bulbs overhead	29	21	15	6	0.2	0.67
d65501	Sweeping outside surfaces	31	20	18	3	0.6	0.67
d6102	Moving light furniture	36	25	12	0	0.2	0.70
d6201	Carrying groceries	35	20	16	8	0.8	0.74
d6402	Cleaning bathroom	33	23	18	7	0.8	0.73
d6402	Making a double bed	37	23	17	5	0.2	0.71
d65502	Tightening nuts and bolts	29	25	18	12	0.4	0.67
d6402	Cleaning windows	32	22	17	14	0.6	0.73
d6600	Helping children dress	30	19	14	23	0.8	0.67
d65505	Walking a leashed dog	24	21	16	27	0.6	0.67
d6600	Helping children wash and dry	30	22	10	26	0.8	0.67
d6503	Washing the car	33	26	20	14	1.6	0.73
d6403	Vacuuming stairs	36	23	19	17	1.4	0.75
d6550	Heavy ironing	34	27	17	17	1.2	0.58
d6201	Carrying bags of fruit	41	22	20	16	1.2	0.76
d6601	Pushing a wheelchair 15 min	37	27	21	18	0.8	0.75
d6601	Helping others up from a seating position	39	29	19	18	0.8	0.74
d65501	Weeding the front yard	43	30	20	16	1.4	0.70
d6405	Depositing two full trash bags	43	26	23	20	1.2	0.75
d65505	Trimming grass using scissors	46	35	26	21	1.4	0.74
d6601	Helping others up from a laying position	50	34	26	21	0.8	0.74
d65505	Pruning hedges	47	36	29	25	1.2	0.73
d65505	Mowing the lawn with a hand mower	45	39	32	25	1.6	0.74
d6102	Redecorating the living room	57	41	32	21	1.2	0.75
d6601	Helping adults in an out of a tub bath	54	40	30	29	0.4	0.72
d65501	Repairing carpentry around the house	47	40	32	36	0.4	0.68
d65501	Painting living room walls	50	40	33	32	2.0	0.72
d65501	Making repairs around the house	52	40	31	32	0.0	0.67
d6102	Assembling a bookcase or wardrobe	53	38	34	30	0.6	0.73
d65505	Spading the garden	57	43	39	31	2.6	0.77
d65501	Hanging wallpaper	54	43	35	40	1.2	0.68
d6102	Carrying moving boxes inside the home	62	44	35	31	1.0	0.75
d65501	Painting the living room	59	42	37	35	2.6	0.74
d6601	Helping a disabled person up two stairs	68	46	37	30	1.0	0.72
d65505	Doing yard work on knees for 1 h	63	48	39	33	2.0	0.75
d6102	Moving heavy furniture	71	49	40	33	0.8	0.75
d6102	Helping carry upstairs a washing machine	70	59	53	43	0.8	0.78
d6102	Helping carry inside heavy furniture	77	59	49	39	1.0	0.78
d6102	Helping someone carry upstairs a sofa	76	61	47	41	1.0	0.78

ICF International Classification of Functioning, Disability and Health; β Partial Credit Model Category intersection parameter; *Effect size* item fit statistic reflecting mean residuals across three score level groups; *H-coefficient Loevinger's* coefficient of scalability, values >0.3 are considered .3 $\leq H_{ij} <$ 0.4 indicates useful but weak *scalability*. 0.4 $\leq H_{ij} <$ 0.5 indicates medium *scalability*. $H_{ij} >$ 0.5 indicates excellent scalability

expected based on their overall level of function, while 3 items all involving repairing vehicles were more difficult for female respondents.

A summary of item characteristics of the final item bank with 41 items is presented in Table 2, organized from least to most difficult to perform activity. The derived IRT theta

scores of the final item bank were transformed into T-scores with a mean of 50 and standard deviation of 10. A T-score of 50 corresponds to the mean level of physical function in the sample and higher scores indicate better function. It can be seen in the table that most of the intersection parameters are localized below the mean, indicating that the items discriminate best between lower levels of physical function. However, the item bank also includes some activities that are well suited to assess respondents with high levels of function (bottom of Table 2).

CAT measurement properties

The correlation between the physical function estimates obtained using the algorithm with 5 items and the estimates for the full item bank obtained in the statistical software used to estimate the item parameters (MIRT) was 0.92 (Table 3) and increased further with the number of administered items. The CAT scores ranged from 22.05 to 79.24. Figure 1 shows that measurement precision of CAT exceeded that of the PF-10, even with only 5 items. Measurement precision was high across the measurement continuum for the CAT with 10 items (SE <7, which roughly corresponds to a classical reliability coefficient of 0.90). By contrast, for the PF-10 this level of precision was observed only for respondents with below average T-scores.

9.4% of respondents had a T-score that exceeded 80 based on the administration of all items. Inspection of response patterns indicated that these respondents scored 0 on all 41 items. These respondents were more frequently male (66.9% vs. 43.7%, $p < 0.01$), significantly younger ($M_{s0} = 38.77$, $M_{s>0} = 50.29$, $p < 0.01$) and significantly less likely to have any chronic (12.8% vs. 44.1%, $p < 0.01$) or rheumatic (1.5% vs. 18.9%, $p < 0.01$) condition than respondents with scores > 0. In terms of ceiling and floor effects, the CAT with 5 items outperformed both the PF-10 and the random administration of 10 REAL items (Table 3). Ceiling effects for the CAT-10 were almost completely resolved for respondents with any rheumatic condition ($n = 2$, 1.1%) and the subpopulation of respondents with any medical condition ($n = 26$, 5.3%). The CAT with 5 items performed equally well as 10 random items in terms of ceiling effects and measurement precision, clearly highlighting the added value of the computerized statistical optimization approach.

Sensitivity

Both CAT-10 and PF-10 discriminated between respondents as indicated by significant F-Tests. T-scores and PF-10 scores increased with the number of chronic conditions and age, and respondents with rheumatic conditions had higher scores than those who did not (Table 4). The relative efficiency of the CAT-10 was greatly superior to the PF-10 for all comparisons.

Respondent feedback

Respondents generally found the questions easy to answer (5-point Likert Mean = 1.5, SD = 1.0) and the wording clear (5-point Likert mean = 4.3, SD = 1.0). 70 patients who found the questions either unclear (n = 83, 7.3%) or difficult to answer (n = 80, 7.1%) motivated their response. 23 respondents reported that a 'non-applicable' response option would have been useful, 11 respondents found some of the questions difficult to answer because they had no experience in doing some of the activities, 5 further respondents directly attributed this difficulty to their gender roles. 7 respondents found the response options difficult to use. The remaining comments pertained to other parts of the survey (e.g., the PF-10 or the question about chronic conditions).

Discussion

In this paper we report on the development, calibration, and initial evaluation of REAL, a new item bank for measuring complex activities of daily living. The results demonstrate that the items fulfill the strong requirements imposed by the

Table 3 Empirical CAT simulation

	T-score, mean (SD)	SEM, mean (SD)	Correlation with full item bank	Floor (%)	Ceiling (%)
Random-10 items	51.89 (14.90)	5.50 (1.80)	0.92	0.4	18.7
CAT-5 items	52.10 (14.92)	5.50 (1.80)	0.92	0.7	15.6
CAT-10 items	54.27 (17.14)	4.00 (1.20)	0.96	0.3	13.1
CAT-15 items	54.43 (17.77	3.60 (130)	0.94	0.1	12.8
PF-10				2.3	38.1

Random-10 administration of 10 random item bank items; *CAT* computerized adaptive test; *PF-10* 36-Item Short-form Health Survey, physical functioning scale; *SEM* standard error of measurement

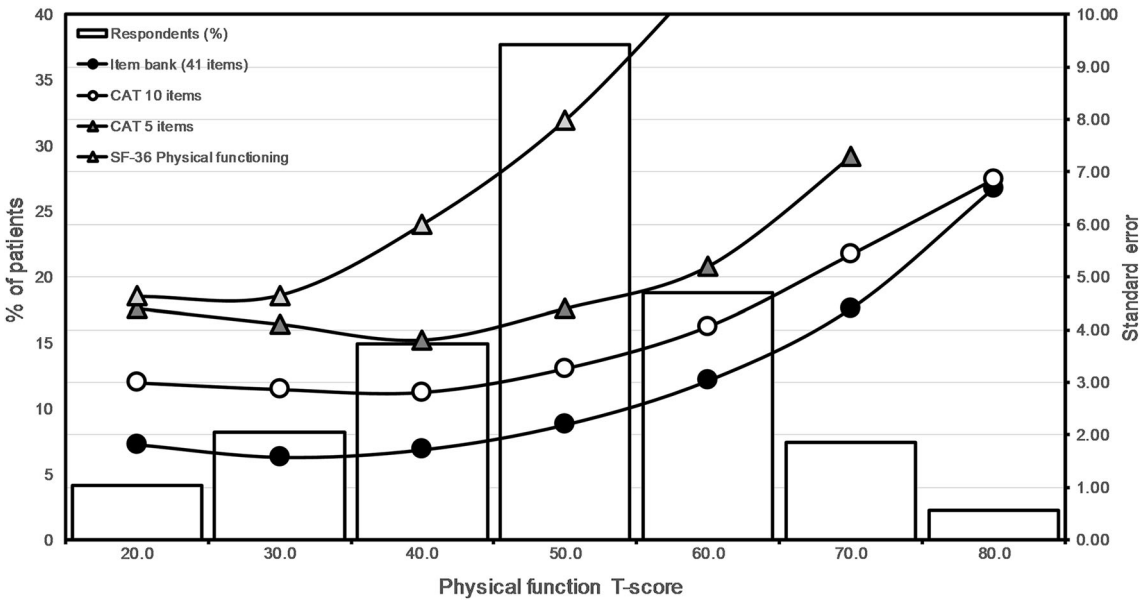

Fig. 1 Measurement precision of a simulated CAT with 5 and 10 items, compared with the PF-10 and the entire item bank

Rasch-type IRT models and support the psychometric quality of the item bank. Moreover, we were able to demonstrate superior measurement performance of CAT versus a traditional pen and paper questionnaire, the PF-10, with particular benefits in higher regions of function.

The current version of the REAL item bank has 41 items that cover most of the second-level codes of the ICF domestic life chapter and were in most cases derived directly from patient responses, which supports the content validity of the item bank [61]. As hypothesized, reliable scores were seen across a wider range of score levels for a CAT with 5 items compared with the PF-10. Although the ceiling effects present in the PF-10 could be reduced by a factor of 3 using CAT with 10 items, some respondents still obtained the highest possible score. In principle, the range of scores that can be reliably measured could be extended further by including even more extreme items (e.g., running 16 miles). However, respondents at the ceiling were typically healthy, younger men, while only slight ceiling effects were observed in the subpopulation with at least one self-reported chronic condition. This suggests that the item bank should yield reliable scores with low ceiling effects in most clinical populations for which physical function levels are likely to be lower than in the subpopulation of respondents with a self-reported chronic condition in the sample studied here.

Although these findings are encouraging, a potential concern with the type of activities included in the item bank is that the frequency with which respondents engage in individual activities may vary with factors such as gender or age. Perceived skill and experience with an activity could therefore contribute to item response behavior and undermine the validity of the scores [37]. However, we used quite conservative definitions of unacceptable DIF and found that no items functioned differently by education or age, while we removed the 10% of evaluated items that were responded to differently by male or female respondents. Since the presence of DIF of the magnitude we tolerated has been shown to have negligible impact on total scores in previous studies, the results of the present study suggest that the validity of scores obtained using the item bank should not be affected much by differential item functioning due to age, sex, or educational level.

The item bank was calibrated using the Rasch-based PCM which is an IRT model that implies that all the information about a person's trait level is provided by the unweighted item scores [62]. The finding that the PCM described the response data well therefore supports the validity of a scoring rule based on summing or averaging of individual items, in cases where respondents are compared using the same set of items (e.g., using a short-form derived from the items in the item bank). Since the patterns of item responses contribute no information to trait level estimates they can also be ignored in other conceivable applications of the item bank such as when developing a crosswalk between a short-form and the T-score metric.

A strength of the study is that we were able to center the scale using a representative sample of the Dutch general population without having to resort to oversampling of respondents with chronic conditions, which would have undermined the validity of norms and might have distorted the item parameters [63]. The well-documented representativeness of the LISS panel as well as the finding that the mean PF-10 score approximates the published Dutch national population norms for this scale [64] supports the

Table 4 Discriminative ability

	Age (years) Mean (SD)			*F* (RE)	Rheumatic condition Mean (SD)		*F* (RE)	Number of conditions Mean (SD)			*F* (RE)
	0–39 ($n = 354$)	40–59 ($n = 362$)	60+ ($n = 410$)		No ($n = 932$)	Yes ($n = 190$)		0 ($n = 631$)	1 ($n = 309$)	2+ ($n = 178$)	
PF-10	90.54 (22.07)	85.42 (24.28)	74.59 (26.30)	42.83 (1.00)	86.35 (24.05)	66.94 (25.08)	101.22 (1.00)	88.18 (23.55)	84.82 (20.50)	61.76 (27.85)	88.89 (1.00)
CAT-10	63.67 (13.47)	57.29 (15.25)	43.75 (15.99)	173.31 (4.12)	57.51 (16.04)	38.61 (13.57)	231.09 (2.20)	59.80 (15.35)	52.16 (15.69)	38.49 (14.92)	137.63 (1.50)

CAT-10 computerized adaptive test with 10 items; *SF-35 PF10* 36-Item Short-form Health Survey, physical functioning scale; *RE* relative efficiency coefficient

use of item bank-derived T-scores for comparing physical function levels to the Dutch adult population. Since items and persons are located on the same scale, a different way to interpret the scores of individuals is in reference to the item steps that are 'dominated' by a person and vice versa. For example, a person with a T-score of 50 has a higher than 50% change to report being able to perform each of the first 25 items of Table 2 without any difficulty.

A limitation of the design of the study is that sparse information was available to explore construct validity in detail and no information was available to study responsiveness. Moreover, the performance of the adaptive testing algorithm was studied using an empirical simulation. Since these item responses were also used to calibrate the item bank, results presented here should be replicated in an actual CAT administration. Work is currently underway to assess the item bank and operational CAT in various rheumatic diseases. An important part of that work will be to investigate the extent to which the item response models obtained here can also be used in populations with different types of disease-related disability and to examine the need for extending the range of the scale to be able to reliably assess respondents with more severe levels of disability than generally observed in the present study. If needed, this could be achieved in several ways. Some of the items that proved to be too easy for respondents in the present study might be calibrated at a later stage. Alternatively, some of the many ADL items that are available in currently validated questionnaires can be included in the item bank.

> In the present study, we have introduced a new item bank focusing on complex activities of daily living that we hope will prove useful for assessing physical function in various populations.

Compliance with ethical standards

Conflict of interest None of the authors declare any conflict of interest.

Ethical approval This study was performed within the larger LISS panel study. According to the Dutch Medical Research Involving Human Subjects Act, such studies do not need approval of an ethical review board. Nevertheless, each respondent provided written consent to be included as a panel member, and data collection abides by the Dutch protection of personal data act.

References

1. Turk, D. C., Dworkin, R. H., Allen, R. R., Bellamy, N., Brandenburg, N., Carr, D. B., … Witter, J. (2003). Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *Pain*, *106*(3), 337–345. doi:10.1016/j.pain.2003.08.001
2. van Tuyl, L. H. D., & Boers, M. (2015). Patient-reported outcomes in core domain sets for rheumatic diseases. *Nature Reviews Rheumatology*. doi:10.1038/nrrheum.2015.116.
3. Salinas, J., Sprinkhuizen, S. M., Ackerson, T., Bernhardt, J., Davie, C., George, M. G., … Schwamm, L. H. (2016). An international standard set of patient-centered outcome measures after stroke. *Stroke*, *47*(1), 180–186. doi:10.1161/STROKEAHA.115.010898
4. Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine, 118*(8), 622–629.
5. Clement, R. C., Welander, A., Stowell, C., Cha, T. D., Chen, J. L., Davies, M., … FRitzell, P. (2015). A proposed set of metrics for standardized outcome reporting in the management of low back pain. *Acta Orthopaedica*, *86*(4), 1–11.
6. Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., … Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation*, *20*(10), 1727–36. doi:10.1007/s11136-011-9903-x
7. McNamara, R. L., Spatz, E. S., Kelley, T. A., Stowell, C. J., Beltrame, J., Heidenreich, P., … Lewin, J. (2015). Standardized outcome measurement for patients with coronary artery disease: consensus from the International Consortium for Health Outcomes Measurement (ICHOM). *Jaha 4*(5), e001767. doi:10.1161/JAHA.115.001767
8. Simonsick, E. M., Newman, A. B., Nevitt, M. C., Kritchevsky, S. B., Ferrucci, L., Guralnik, J. M., … Health ABC Study Group. (2001). Measuring higher level physical function in well-functioning older adults: expanding familiar approaches in the Health ABC study. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *56*(10), M644-9.
9. Collins, K., Rooney, B. L., Smalley, K. J., & Havens, S. (2004). Functional fitness, disease and independence in community-dwelling older adults in western Wisconsin. *WMJ: official publication of the State Medical Society of Wisconsin, 103*(1), 42–48.
10. Stucki, G., Stucki, S., Brühlmann, P., & Michel, B. A. (1995). Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. *Annals of the Rheumatic Diseases, 54*(6), 461–465.
11. Dunbar, M. J., Robertsson, O., Ryd, L., & Lidgren, L. (2001). Appropriate questionnaires for knee arthroplasty. Results of a survey of 3600 patients from The Swedish Knee Arthroplasty Registry. *The Journal of Bone and Joint Surgery. British Volume*, *83*(3), 339–44.
12. Hsueh, I. P., Lee, M. M., & Hsieh, C. L. (2001). Psychometric characteristics of the Barthel activities of daily living index in stroke patients. *Journal of the Formosan Medical Association = Taiwan yi zhi, 100*(8), 526–532.
13. Fries, J., Rose, M., & Krishnan, E. (2011). The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. *The Journal of Rheumatology, 38*(8), 1759–1764. doi:10.3899/jrheum.110402.
14. Wolfe, F., Michaud, K., & Pincus, T. (2004). Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. *Arthritis and Rheumatism, 50*(10), 3296–3305. doi:10.1002/art.20549.
15. Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 6*(6), 595–600.
16. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory. Measurement methods for the social sciences series, Vol. 2*. Newbury Park, CA: Sage.
17. Ayala, R. De. (2013). The theory and practice of item response theory. New York: Guilford Press
18. van der Linden, W., & Glas, C. A. W. (2010). Elements of adaptive testing. *Elements*. doi:10.1007/978-0-387-85461-8.
19. Oude Voshaar, M. A. H., Ten Klooster, P. M., Glas, C. A. W., Vonkeman, H. E., Krishnan, E., & van de Laar, M. A. F. J. (2014). Relative performance of commonly used physical function questionnaires in rheumatoid arthritis and a patient-reported outcomes measurement information system computerized adaptive test. *Arthritis & Rheumatology, 66*(10), 2900–2908. doi:10.1002/art.38759.
20. Weisscher, N., Post, B., de Haan, R. J., Glas, C. A. W., Speelman, J. D., & Vermeulen, M. (2007). The AMC Linear Disability Score in patients with newly diagnosed Parkinson disease. *Neurology, 69*(23), 2155–2161. doi:10.1212/01.wnl.0000295666.30948.9d.
21. Li, C.-Y., Romero, S., Bonilha, H. S., Simpson, K. N., Simpson, A. N., Hong, I., et al. (2016). Linking existing instruments to develop an activity of daily living item bank. *Evaluation and the Health Professions*. doi:10.1177/0163278716676873.
22. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526. doi:10.1016/j.jclinepi.2013.10.024.
23. Holman, R., Weisscher, N., Glas, C. A. W., Dijkgraaf, M. G. W., Vermeulen, M., de Haan, R. J., et al. (2005). The Academic Medical Center Linear Disability Score (ALDS) item bank: item response theory analysis in a mixed patient population. *Health and Quality of Life Outcomes, 3*(1), 83. doi:10.1186/1477-7525-3-83.
24. Fries, J. F., Lingala, B., Siemons, L., Glas, C. A. W., Cella, D., Hussain, Y. N., … Krishnan, E. (2014). Extending the floor and the ceiling for assessment of physical function. *Arthritis & rheumatology (Hoboken, N.J.)*, *66*(5), 1378–87. doi:10.1002/art.38342
25. Hung, M., Clegg, D. O., Greene, T., & Saltzman, C. L. (2011). Evaluation of the PROMIS physical function item bank in orthopaedic patients. *Journal of Orthopaedic Research, 29*(6), 947–953. doi:10.1002/jor.21308.
26. Petersen, M. A., Groenvold, M., Aaronson, N. K., Chie, W.-C., Conroy, T., Costantini, A., … Young, T. (2011). Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *20*(4), 479–90. doi:10.1007/s11136-010-9770-x
27. Bruce, B., Fries, J. F., Ambrosini, D., Lingala, B., Gandek, B., Rose, M., et al. (2009). Better assessment of physical function: item improvement is neglected but essential. *Arthritis Research & Therapy, 11*(6), R191. doi:10.1186/ar2890.
28. Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care, 19*(8), 787–805.
29. Collin, C., Wade, D. T., Davies, S., & Horne, V. (1988). The Barthel ADL Index: a reliability study. *International Disability Studies, 10*(2), 61–63.

30. Fries, J. F., Spitz, P., Kraines, R. G., & Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis and Rheumatism, 23*(2), 137–145.
31. Kempen, G. I., & Suurmeijer, T. P. (1990). The development of a hierarchical polychotomous ADL-IADL scale for noninstitutionalized elders. *The Gerontologist, 30*(4), 497–502.
32. Spector, W. D., Katz, S., Murphy, J. B., & Fulton, J. P. (1987). The hierarchical relationship between activities of daily living and instrumental activities of daily living. *Journal of Chronic Diseases, 40*(6), 481–489.
33. Njegovan, V., Hing, M. M., Mitchell, S. L., & Molnar, F. J. (2001). The hierarchy of functional loss associated with cognitive decline in older persons. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences, 56*(10), M638–M643.
34. Kempen, G. I., Miedema, I., Ormel, J., & Molenaar, W. (1996). The assessment of disability with the Groningen Activity Restriction Scale. Conceptual framework and psychometric properties. *Social Science & Medicine (1982), 43*(11), 1601–1610.
35. Ng, T.-P., Niti, M., Chiam, P.-C., & Kua, E.-H. (2006). Physical and cognitive domains of the instrumental activities of daily living: validation in a multiethnic population of Asian older adults. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences, 61*(7), 726–735.
36. Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist, 9*(3), 179–186.
37. Reuben, D. B., & Solomon, D. H. (1989). Assessment in geriatrics. Of caveats and names. *Journal of the American Geriatrics Society, 37*(6), 570–572.
38. Lindeboom, R., Vermeulen, M., Holman, R., & De Haan, R. J. (2003). Activities of daily living instruments: Optimizing scales for neurologic assessments. *Neurology, 60*(5), 738–742.
39. Weigl, M., Cieza, A., Harder, M., Geyh, S., Amann, E., Kostanjsek, N., et al. (2003). Linking osteoarthritis-specific health-status measures to the International Classification of Functioning, Disability, and Health (ICF). *Osteoarthritis and Cartilage, 11*(7), 519–523.
40. Oude Voshaar, M. A. H., ten Klooster, P. M., Taal, E., & van de Laar, M. A. F. J. (2011). Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. *Health and Quality of Life Outcomes, 9,* 99. doi:10.1186/1477-7525-9-99.
41. Oude Voshaar, M. A. H., ten Klooster, P. M., Glas, C. A. W., Vonkeman, H. E., Taal, E., Krishnan, E., … van de Laar, M. A. F. J. (2015). Validity and measurement precision of the PROMIS physical function item bank and a content validity–driven 20-item short form in rheumatoid arthritis compared with traditional measures. *Rheumatology*, *54*(12), kev265. doi:10.1093/rheumatology/kev265
42. Forget, N., & Higgins, J. (2014). Comparison of generic patient-reported outcome measures used with upper extremity musculoskeletal disorders: Linking process using the International Classification of Functioning, Disability, and Health (ICF). *Journal of Rehabilitation Medicine, 46*(4), 327–334. doi:10.2340/16501977-1784.
43. Stier-Jarmer, M., Cieza, A., Borchers, M., Stucki, G., & World Health Organization. (2009). How to apply the ICF and ICF core sets for low back pain. *The Clinical Journal of Pain, 25*(1), 29–38. doi:10.1097/AJP.0b013e31817bcc78.
44. Cieza, A., Geyh, S., Chatterji, S., Kostanjsek, N., Ustün, B., & Stucki, G. (2005). ICF linking rules: an update based on lessons learned. *Journal of Rehabilitation Medicine, 37*(4), 212–218. doi:10.1080/16501970510040263.
45. Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: how the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology*, *109*(1), 56–61.
46. Brüggen, E., van den Brakel, J., & Krosnick, J. (2016). Establishing the accuracy of online panels for survey research. *forthcoming paper*.
47. Scherpenzeel, A., & Bethlehem, J. (2011). How representative are online panels? Problems of coverage and selection and possible solutions. In M. Das, P. Ester, L. Kaczmirek, & P. Mohler (Eds.), *Social research and the internet: Advances in applied methods and new research strategies* (pp. 105–132). New York: Routledge Academic.
48. Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*(6), 473–483.
49. Yu, C. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.
50. Van der Ark, L. A., & Andries, L. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1–19.
51. van Schuur, W. (2011). *Ordinal item response theory: Mokken scale analysis.* (169th ed.). London: Sage.
52. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*.
53. Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273–294. doi:10.1007/BF02294296.
54. Glas, C. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*, 647–667.
55. International Standard Classification of Education - Google Scholar. (n.d.). Retrieved February 19, 2017, from https://scholar.google.com/scholar?hl=nl&q=International+Standard+Classification+of+Education&btnG=&lr=
56. Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.
57. Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.
58. Glas, C. (2010). *Preliminary manual of the software program Multidimensional Item Response Theory (MIRT)*. Enschede: University of Twente.
59. Penfield, R. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*(1), 1–20.
60. Fayers, P., & Machin, D. (2013). Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. Chichester: Wiley
61. Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., … de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42. doi:10.1016/j.jclinepi.2006.03.012
62. Sijtsma, K., & Hemker, B. T. (2017). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics, 25*(4), 391–415.
63. Smits, N. (2016). On the effect of adding clinical samples to validation studies of patient-reported outcome item banks: a simulation study. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 25*(7), 1635–1644. doi:10.1007/s11136-015-1199-9.
64. Aaronson, N. K., Muller, M., Cohen, P. D., Essink-Bot, M. L., Fekkes, M., Sanderman, R., … Verrips, E. (1998). Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *Journal of clinical epidemiology*, *51*(11), 1055–1068.