

Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension

Morten Aa. Petersen · Mogens Groenvold · Neil K. Aaronson · Wei-Chu Chie · Thierry Conroy · Anna Costantini · Peter Fayers · Jorunn Helbostad · Bernhard Holzner · Stein Kaasa · Susanne Singer · Galina Velikova · Teresa Young

Accepted: 4 October 2010 / Published online: 23 October 2010
© Springer Science+Business Media B.V. 2010

Abstract

Purpose Computerized adaptive test (CAT) methods, based on item response theory (IRT), enable a patient-reported outcome instrument to be adapted to the individual patient while maintaining direct comparability of scores. The EORTC Quality of Life Group is developing a

CAT version of the widely used EORTC QLQ-C30. We present the development and psychometric validation of the item pool for the first of the scales, physical functioning (PF).

Methods Initial developments (including literature search and patient and expert evaluations) resulted in 56 candidate items. Responses to these items were collected from 1,176 patients with cancer from Denmark, France, Germany,

This study is conducted on behalf of the EORTC Quality of Life Group.

M. Aa. Petersen (✉) · M. Groenvold
The Research Unit, Department of Palliative Medicine,
Bispebjerg Hospital, Bispebjerg bakke 23,
2400 Copenhagen NV, Denmark
e-mail: mpet0009@bbh.regionh.dk

M. Groenvold
Institute of Public Health, University of Copenhagen,
Copenhagen, Denmark

N. K. Aaronson
Division of Psychosocial Research and Epidemiology,
The Netherlands Cancer Institute, Amsterdam, The Netherlands

W.-C. Chie
Graduate Institute of Preventive Medicine and Department
of Public Health, College of Public Health, National Taiwan
University, Taipei, Taiwan

T. Conroy
Medical Oncology Department, Centre Alexis Vautrin,
Vandoeuvre-lès-Nancy, France

A. Costantini
Psychoncology Unit, Sant'Andrea Hospital, 2nd Faculty
of Medicine, Sapienza University of Rome, Rome, Italy

P. Fayers
Department of Public Health, University of Aberdeen, Aberdeen,
UK

P. Fayers
Department of Cancer Research and Molecular Medicine,
Norwegian University of Science and Technology,
Trondheim, Norway

J. Helbostad
Department of Neuroscience, Norwegian University of Science
and Technology and St. Olav University Hospital,
Trondheim, Norway

B. Holzner
Department of Psychiatry and Psychotherapy,
Medical University Innsbruck, Innsbruck, Austria

S. Kaasa
Palliative Medicine Unit, University Hospital of Trondheim,
Trondheim, Norway

S. Singer
Department Medical Psychology and Medical Sociology,
University of Leipzig, Leipzig, Germany

G. Velikova
Cancer Research UK Centre, University of Leeds, Leeds, UK

T. Young
Lynda Jackson Macmillan Centre, Mount Vernon Cancer Centre,
Northwood, Middx, UK

Italy, Taiwan, and the United Kingdom. The items were evaluated with regard to psychometric properties.

Results Evaluations showed that 31 of the items could be included in a unidimensional IRT model with acceptable fit and good content coverage, although the pool may lack items at the upper extreme (good PF). There were several findings of significant differential item functioning (DIF). However, the DIF findings appeared to have little impact on the PF estimation.

Conclusions We have established an item pool for CAT measurement of PF and believe that this CAT instrument will clearly improve the EORTC measurement of PF.

Keywords Computerized adaptive test · EORTC QLQ-C30 · Item banking · Item response theory · Physical functioning · Quality of life

Introduction

The European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) [1] is one of the most widely used health-related quality of life (HRQOL) questionnaires in cancer research [2, 3]. The questionnaire consists of 30 items measuring 15 aspects of HRQOL: five functional measures, nine symptom measures, and one measure of overall health/quality of life [4]. The EORTC QLQ-C30 is a traditional, standardized questionnaire, i.e., all patients are asked the same 30 questions. This ensures comparability of scores across patients, but also means the questionnaire may not always be optimal: some patients may answer irrelevant questions, and some HRQOL domains may be measured with less precision than is desired.

Therefore, the EORTC Quality of Life Group (QLG) initiated a project to develop a computerized adaptive test (CAT) [5] version of the QLQ-C30 [6]. CAT is a computer-based, interactive method for constructing individualized instruments. Based on a patient's previous responses, a computer program evaluates which item should be asked next to obtain maximal information. Additional items are administered until a prespecified level of precision has been reached or until a predefined number of items has been administered. Basing the CAT on item response theory (IRT) methods [7, 8] ensures that scores are directly comparable across patients, even though they answer different subsets of questions. CAT measurement has several advantages over traditional questionnaires, perhaps most importantly that fewer items are required to obtain the same level of precision [5].

The aim is to develop CAT versions of the HRQOL domains included in the QLQ-C30, which are more efficient, precise and flexible than the original instrument. This

requires the development of additional items supplementing the original QLQ-C30 items. Our development of these so-called "item pools" involves the following steps for each HRQOL domain: (1) a literature search to obtain a clear understanding of the domain and to identify existing items measuring that domain; (2) formulating new items to assess the relevant aspects of the domain, using the same item format as that employed with the QLQ-C30; (3) conducting patient interviews to pre-test the items; and (4) collecting data and conducting psychometric analyses for the final selection of items and calibration of the IRT model forming the basis of the CAT. Details of the project and the methods of steps 1–3 have previously been described [6].

In developing the CAT measure for physical functioning (PF), we identified 975 PF items from the literature. Of these, 407 items measured one of the aspects of PF included in the QLQ-C30: lifting and carrying objects, walking and moving, mobility (unspecified), and self-care. These items were used as inspiration to develop 86 new items, which were reduced to 66 based on reviews by the EORTC CAT-project group. Based on expert and patient evaluations, several items were revised and the list was further reduced to 51 items [6].

This paper reports on step 4 in the development for PF:

1. Evaluation of the psychometric characteristics of the candidate items, particularly their fit to a unidimensional IRT model measuring "QLQ-C30" PF.
2. Final selection of items for the PF item pool and calibration of the IRT model for the CAT.

Methods

Sample

Precise guidelines for sample size requirements for IRT/CAT analyses are not available, but simulations and practical experience suggest that 500–1,000 observations suffice for precise calibration [9–11]. The precision for a given sample size will of course depend on many factors, such as distribution of item responses and the composition of patients in the sample. Therefore, we aimed for at least 1,000 responses. As the CAT is intended for use internationally and across a wide range of patients with cancer, we accrued a heterogeneous patient sample including different age groups, diagnoses, etc. Patients were recruited from Denmark, France, Germany, Italy, Taiwan, and the United Kingdom. Eligibility requirements included having had a cancer diagnosis, age greater than 18 years, and being physically and mentally competent to complete the questionnaire.

The study was approved by the local ethics committees of the participating countries.

Questionnaire

The questionnaire included: (1) the 51 candidate PF items [6] plus the five original PF items from the QLQ-C30, in all 56 PF items, in random order; (2) the remainder of the QLQ-C30; and (3) debriefing items to determine whether the patients found any of the questions problematic. We also collected information about the patients' age, gender, cancer site, cancer stage, current treatment, level of education, employment status, and marital/cohabiting status.

The focus here is on the PF items. All items employ a 4-point response scale: “not at all”, “a little”, “quite a bit”, and “very much”. As in the QLQ-C30 PF scale, the items do not refer to a specific time frame, but ask about performing a task generally: “Do you have any trouble...”/ “Do you need help...”.

Analysis plan

The evaluation and final selection of items for the PF CAT was organized into six steps:

Descriptive and basic statistical analyses

Calculation of proportion missing responses, item frequency distributions, means and standard deviations, and correlations with the original 5-item QLQ-C30 PF sum scale.

Patient comments and evaluations of the items

Based on patients' qualitative comments and responses to the debriefing questions, we evaluated whether some of the items should be deleted because a substantial number of patients found them problematic. If only a few patients found an otherwise well-functioning item problematic, we would not remove it, since this might result in impaired measurement properties for the large majority.

Evaluation of dimensionality and local dependence

Our aim was a unidimensional CAT measuring the same aspects of PF as the QLQ-C30 PF scale. To investigate the dimensionality of the item set, we used factor analysis methods for ordinal categorical data [12]. First, we explored the dimensionality based on eigenvalues using scree plots [13] and proportion of variance explained. Next, we searched for a unidimensional model including the five QLQ-C30 PF items and as many of the candidate items as

possible. The items should have good content coverage (with regard to both aspects and levels of PF) and be essentially unidimensional. The following fit criteria were used to indicate reasonable fit: the root mean square error of approximation (RMSEA), the Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI). Rules of thumb are that $RMSEA < 0.10$ and TLI and $CFI > 0.90$ indicate reasonable (not poor) fit [14, 15]. Coverage was evaluated by inspecting coverage of the different PF aspects, the test information function, and the variability of item difficulties.

A requirement of standard IRT models is that the items are locally independent, i.e., item responses are independent when controlling for the overall PF score. We investigated this using the residual correlations from the final factor analysis model. It has been proposed that residual correlations > 0.20 [16] or > 0.25 [17] indicate local dependence.

Calibration of IRT model and evaluation of item fit

Standard IRT models assume monotonicity: the cumulative probability of choosing a given response category or a higher category does not decrease with increasing IRT scores. We investigated each item for monotonicity by inspecting the average item score in relation to the rest score (the sum score of all items except the item in question). If an item complies with monotonicity, the average item score is not expected to decrease for increasing values of the rest score [18].

We used the generalized partial credit model (GPCM) as the IRT model forming the basis for the PF CAT [19]. In the GPCM, each item has a slope parameter describing the item's ability to discriminate between subjects with different levels of PF, and a set of threshold parameters describing how likely it is to report problems on the item. An advantage of the GPCM is that it is a generalization of other well-known item models. If all items have the same slope, the model reduces to the partial credit model, which belongs to the family of Rasch models [20]. To evaluate the effect of model choice on the item fit, we planned also to calibrate the graded response model [21], which has the same number of item parameters as the GPCM and often gives trace lines that are very similar to the GPCM trace lines [22].

We initially used Parscale for estimating the IRT model [11]. However, the estimation failed because of convergence problems. Therefore, we used the IRT Command Language (ICL) program [23]. ICL computes marginal maximum likelihood estimates of the item parameters using the expectation–maximization (EM) algorithm [24, 25]. ICL can only estimate the GPCM. Therefore, in the following, we restricted our model search and evaluations to this model. The estimation procedure of ICL

seemed more robust, but ICL does not include tests of model fit. Therefore, the parameter estimates from ICL were imported into Parscale and examined for item fit using Muraki's test [11, 19]. Because of multiple testing and a large sample, we used $P < 0.001$ as indicating misfit. Furthermore, we calculated the average difference between expected and observed item responses (bias) and the root mean square error (RMSE) of expected and observed item responses. Finally, we calculated the information function for the selected set of items to evaluate whether the items seemed to have acceptable measurement precision across the continuum.

Test for differential item functioning (DIF)

The IRT model assumes that patients with the same level of PF have the same probability of responding to an item in a certain way, regardless of group membership. If this is not the case, the item is said to show differential item functioning (DIF).

We explored for DIF using ordinal logistic regression [26, 27]. For each item, the item response was regressed on the estimated PF score, the group variable and the interaction of the two. A significant interaction indicates non-uniform DIF (the DIF differs across the PF scores), while a significant group parameter indicates uniform DIF (the same DIF across the PF scores). The PF score was estimated from the final IRT model calibrated in step 4.

We explored for DIF with regard to gender, age (<40, 40–49, 50–59, 60–69, ≥ 70 years), country (Denmark, France, Germany, Italy, Taiwan, United Kingdom), cancer site (breast, gastrointestinal, urogenital, gynaecological, head and neck, lung, other), cancer stage (stage I–II vs. III–IV), current treatment (chemotherapy, other treatment, no current treatment), education (0–10, 11–13, 14–16, >16 years of education), work (working, retired, other), and cohabitation (living with a partner vs. living alone).

Because of a large sample and multiple testing, we used a significance level of $P < 0.001$. Significant and potentially relevant uniform DIF was implied if $P < 0.001$ and the regression coefficient for the group variable was >0.64 (for group variables with more than two groups, at least two groups' coefficients should differ >0.64) [27, 28]. As a criterion for significant and potentially relevant non-uniform DIF, we used $P < 0.001$ and a change in the Nagelkerke R^2 -coefficient when deleting the interaction of at least 0.035 [29, 30]. When an item showed significant DIF with regard to two or more group variables, the significant group variables were entered in a multiple logistic regression model to examine whether the findings seemed to reflect genuine DIF, i.e., these analyses were intended to eliminate false positive DIF findings caused by

confounding of the group variables. Only group variables showing significant DIF in these “multiple variable” DIF analyses are reported.

Even though DIF may have significant impact at the item level, the DIF may have little impact at the scale level [31]. For example, DIF in one direction for an item may be neutralized by DIF in the opposite direction for other items (also see Discussion). To evaluate the practical importance of the DIF findings for the estimation of PF, we estimated a new IRT model that calibrated the item parameters independently in the DIF groups for the item of focus, fixing the remainder of the items' parameters to the estimates from step 4. We then compared the PF scores obtained with this model accounting for DIF with the scores obtained with the model from step 4, which ignores any possible DIF. If the PF scores obtained with the two models were similar, the DIF in the item was considered not to have practical importance for PF estimation. We calculated the correlation and difference between the two PF estimations and the percentage of patients with differences greater than the median standard error for the PF estimates. This has been suggested as a criterion for “salient scale-level differential functioning” [32].

The evaluations of DIF were performed using SAS v. 9.1.3 [33].

Evaluation of discarded items and final IRT calibration

To ensure that an item had not erroneously been discarded in one of the steps above, we added the discarded items one at a time to the list of items obtained after step 5 and evaluated whether the item still showed misfit. Items that had erroneously been discarded were included again.

Results

Sample

In total, we collected responses from 1,176 patients with cancer. Patient characteristics are reported in Table 1.

Table 2 shows basic descriptive information for the 56 PF items. Response rates were generally high (97.2–99.4%). Most items had skewed response distributions. Particularly, the items about self-care had highly skewed response distributions (for all except one self-care item, at least 90% of the patients responded “not at all”). Polychoric correlations for the 51 new items with the sum scale based on the five QLQ-C30 PF items ranged from 0.64 to 0.93. In comparison, the five QLQ-C30 PF items correlated 0.63–0.85 with the sum scale corrected for overlap (i.e., excluding the item in question).

Table 1 Sociodemographic and clinical characteristics of the study sample ($N = 1176$)

	<i>N</i> /mean
Age (mean years)	58 (range 18–91)
Gender	
Male	524 (45%)
Female	648 (55%)
Country	
Denmark	412 (35%)
France	314 (27%)
Germany	163 (14%)
Italy	87 (7%)
Taiwan	100 (9%)
UK	100 (9%)
Education (years)	
0–10	315 (27%)
11–13	265 (23%)
14–16	280 (24%)
>16	281 (24%)
Work	
Working	389 (33%)
Retired	557 (47%)
Other	212 (18%)
Cohabitation	
Living with a partner	844 (72%)
Living alone	305 (26%)
Cancer stage	
I–II	399 (34%)
III–IV	583 (50%)
Cancer site	
Breast	150 (13%)
Gastrointestinal	135 (11%)
Gynaecological	180 (15%)
Head and neck	163 (14%)
Lung	52 (4%)
Urogenital	181 (15%)
Other	124 (11%)
Current treatment	
Chemotherapy	443 (38%)
Other treatment	97 (8%)
No current treatment	605 (52%)

Patient comments and evaluations of the items

In all, 148 patients (12.6%) reported one or more items as being problematic. Item 17 “Do you have any trouble participating in strenuous sports, like running 10 km, cycling 25 km, or a similar activity?” was most often perceived as problematic: 35 patients (3.0%) found the item difficult to answer and 15 (1.3%) found it confusing.

Twenty-one patients (1.8%) found items 34, 35, and 46 difficult to answer, and 12–15 patients (1.0–1.3%) found items 16, 27, 30 (from the QLQ-C30), 45, 47 (from the QLQ-C30), and 48 difficult to answer. All other items were rated as problematic by less than 10 patients (<1%). Except for four items rated as upsetting or intrusive by one patient, none of the patients found any of the items upsetting or intrusive. Overall, very few patients found the items problematic, and when reporting problems, it was typically that an item was difficult to answer. We did not remove any items based solely on these patient comments.

Evaluation of dimensionality and local dependence

Exploratory eigenvalues analysis showed that about two-thirds of the total variation was explained by the first factor (Table 3). Five factors had eigenvalues above 1, but only two factors explained more than 5% of the variation, and the scree plot also suggested two factors (details not shown). Including all 56 items in a unidimensional model resulted in $RMSEA = 0.17$, $CFI = 0.89$, $TLI = 0.97$. Together these analyses suggest that it might not be sensible to include all 56 items in a unidimensional model. However, since there was one clearly dominating factor, it seemed plausible that we could find a unidimensional model, including the QLQ-C30 PF items and a reasonable number of the candidate items. Using an exploratory factor analysis approach to divide the items into two or more factors did not result in meaningful factors: each factor was a mix of items from the different PF aspects and of items of different “difficulty”, e.g., some walking items loaded on one factor, while other loaded on another. Therefore, as the next step, we investigated the fit of unidimensional models consisting of the QLQ-C30 PF items and various subsets of the candidate items. We started with the items that seemed to have the best fit based on the results from the previous steps, evaluated these one at a time (in the model including the QLQ-C30 items), and included them if it did not result in poor model fit. At the same time, we tried to include items with an eye towards maintaining adequate content coverage (with regard to both aspects and levels of PF). All candidate items were evaluated in these analyses.

In this way, it was possible to include 34 of the 56 items. The fit indices of this model were $RMSEA = 0.09$, $CFI = 0.94$, and $TLI = 0.98$ indicating acceptable to good fit. If we added any additional items, at least one of the fit indices indicated poor fit. A scree plot for these 34 items also indicated that one factor would be sufficient (details not shown).

Inspection of residual correlations for the 34 items revealed that, of the 561 possible correlations, seven (1.2%) were >0.2 and three (0.5%) were >0.25 (details not shown). Since no clear pattern was observed in these

Table 2 Descriptive statistics for the 56 candidate items for the PF item pool

Item no, abbreviated item text	Response rate (%)	Mean (SD)	% “Not at all” (ceiling)	Correlation ^b
Item 1, lifting a full cup or glass to your mouth	99.4	2.93 (0.34)	94.6	0.65
Item 2, help to walk about indoors	99.4	2.88 (0.48)	92.6	0.81
Item 3, walking 100 m	98.6	2.70 (0.71)	81.3	0.88
Item 4, help caring for your feet	99.4	2.60 (0.86)	78.0	0.65
Item 5, walking around indoors	99.1	2.86 (0.49)	90.7	0.77
Item 6, walking up a flight of stairs	99.1	2.58 (0.79)	72.4	0.83
Item 7, help eating, dressing, washing ^a	99.3	2.91 (0.41)	94.3	0.78
Item 8, help with grooming	99.3	2.93 (0.34)	95.0	0.83
Item 9, walking down a flight of stairs	99.3	2.74 (0.64)	83.1	0.82
Item 10, walking over uneven ground	99.1	2.64 (0.72)	75.2	0.81
Item 11, walking for 30 min	98.9	2.36 (0.98)	63.3	0.88
Item 12, walking 100 m on level ground	98.8	2.68 (0.74)	81.5	0.88
Item 13, help putting on or taking off trousers	99.2	2.91 (0.40)	94.1	0.80
Item 14, help taking a shower	99.1	2.86 (0.53)	91.8	0.82
Item 15, lifting a box weighing 10 kg	99.3	2.13 (1.08)	52.2	0.81
Item 16, hiking 3 km	98.6	1.88 (1.18)	43.1	0.89
Item 17, participating in strenuous sports	97.6	0.95 (1.09)	14.0	0.78
Item 18, help dressing	99.4	2.91 (0.39)	94.4	0.81
Item 19, help putting on your shoes	99.3	2.88 (0.45)	92.2	0.80
Item 20, trouble taking a long, brisk walk	98.6	1.91 (1.13)	41.9	0.88
Item 21, walking outdoors on flat ground	98.4	2.65 (0.76)	78.1	0.85
Item 22, help brushing your teeth	98.6	2.97 (0.21)	98.0	0.73
Item 23, walking 500 m	98.4	2.48 (0.93)	71.1	0.87
Item 24, climbing three flights of stairs	98.2	2.01 (1.10)	45.0	0.85
Item 25, walking a few steps	98.5	2.91 (0.39)	93.5	0.81
Item 26, stay in bed or a chair during the day ^a	98.5	2.59 (0.73)	71.0	0.63
Item 27, lifting a box weighing 10 kg and carrying it	97.9	1.93 (1.15)	44.3	0.83
Item 28, help pulling on a sweater	98.6	2.93 (0.36)	95.1	0.77
Item 29, help putting on a shirt	98.5	2.93 (0.35)	95.8	0.79
Item 30, doing strenuous activities ^a	98.9	1.95 (1.07)	40.0	0.77
Item 31, walking 100 m carrying a heavy shopping bag	98.4	1.86 (1.12)	38.7	0.88
Item 32, getting down an object weighing 2 kg	98.9	2.50 (0.88)	70.0	0.79
Item 33, bending over	98.5	2.62 (0.75)	74.6	0.79
Item 34, running 1 km	97.5	1.25 (1.18)	21.4	0.82
Item 35, running fast	97.4	1.15 (1.16)	18.7	0.81
Item 36, walking 10 m inside	98.4	2.84 (0.53)	90.0	0.78
Item 37, taking a short walk ^a	99.0	2.78 (0.59)	85.0	0.85
Item 38, taking a brisk walk	99.0	2.22 (0.99)	52.7	0.89
Item 39, walking without losing your balance	98.9	2.72 (0.64)	80.9	0.75
Item 40, walking 1 km	98.3	2.24 (1.08)	60.3	0.89
Item 41, carrying a heavy bag upstairs	98.3	1.90 (1.09)	38.4	0.89
Item 42, help washing your face and hands	99.1	2.97 (0.22)	98.1	0.85
Item 43, help to go to the lavatory	98.7	2.96 (0.26)	97.7	0.93
Item 44, walking 10 m outside	98.5	2.84 (0.50)	89.0	0.78
Item 45, running 100 m	97.2	1.61 (1.21)	33.6	0.83
Item 46, taking a long walk carrying a heavy pack	97.4	1.43 (1.20)	26.8	0.89
Item 47, taking a long walk ^a	97.8	1.95 (1.17)	47.3	0.79
Item 48, running up three flights of stairs	97.9	1.30 (1.21)	24.0	0.82

Table 2 continued

Item no, abbreviated item text	Response rate (%)	Mean (SD)	% “Not at all” (ceiling)	Correlation ^b
Item 49, carrying something weighing 5 kg	98.5	2.11 (1.09)	51.9	0.83
Item 50, help to walk about outside	98.5	2.78 (0.66)	87.6	0.83
Item 51, help washing and drying whole body	98.7	2.89 (0.45)	92.7	0.77
Item 52, help undressing	99.0	2.93 (0.34)	95.6	0.84
Item 53, carrying something in both hands while climbing stairs	98.8	2.03 (1.10)	47.3	0.85
Item 54, help eating	99.1	2.96 (0.26)	97.5	0.64
Item 55, lifting a full teapot/coffeepot	98.8	2.86 (0.47)	89.6	0.74
Item 56, running a short distance	98.5	1.86 (1.18)	42.5	0.86

The items are scored: 0 = “very much”, 1 = “quite a bit”, 2 = “a little”, and 3 = “not at all”

^a EORTC QLQ-C30 PF item

^b Polychoric correlation with the sum scale based on the five QLQ-C30 PF items. For each of the five QLQ-C30 items, the correlation is with the sum scale of the other four C30 PF items

Table 3 Eigenvalue analysis of the initial 56 candidate items and for the 31 items in the final PF pool, respectively

Factors	F1	F2	F3	F4	F5	F6	F7
Initial 56 items							
Eigenvalues	37.4	5.9	2.1	1.3	1.0	0.8	0.6
Difference in eigenvalues	31.5	3.9	0.7	0.3	0.2	0.2	0.0
Explained variation (%)	66.8	10.6	3.7	2.4	1.8	1.5	1.1
Cumulated explained variation (%)	66.8	77.4	81.1	83.5	85.3	86.8	87.9
Final 31 items							
Eigenvalues	22.1	2.0	1.0	0.7	0.6	0.5	0.5
Difference in eigenvalues	20.1	1.0	0.3	0.1	0.1	0.0	0.0
Explained variation (%)	71.3	6.4	3.3	2.4	2.1	1.8	1.6
Cumulated explained variation (%)	71.3	77.7	81.0	83.4	85.5	87.2	88.8

correlations (i.e., they may just be random findings), all 34 items were retained in a unidimensional PF model.

Calibration of the IRT model and evaluation of item fit

There were no indications of violations of the assumption of monotonicity for the 34 items (details not shown). Therefore, we calibrated a GPCM to all 34 items. Because the highly skewed distributions of some items might make calibration problematic and to obtain more robust parameter estimates, we collapsed (combined) response categories with less than 10 responses. This means that some items only have two or three effective categories. Fewer categories mean that less information can be obtained with an item, but we judged this better than to risk getting imprecise parameter estimates or discarding such items.

Further, some items had disordered threshold parameter estimates (lower threshold for higher category). Technically,

this might not be a problem, but it might reflect that some items have lacked responses in the upper categories to obtain robust parameter estimates or that the patients have had difficulties distinguishing between/correctly using the response categories for some items. Hence, to avoid such potential problems, such categories were collapsed (see Table 4).

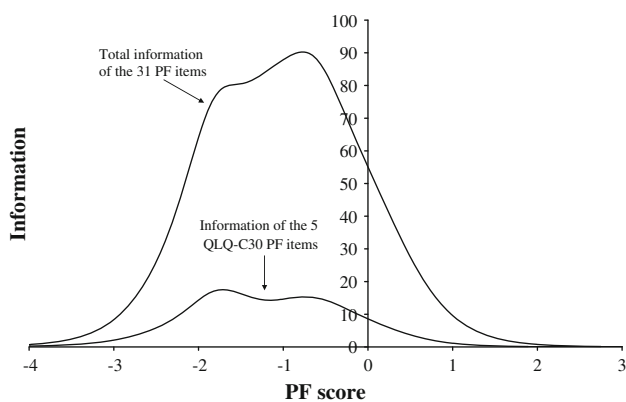
Based on the evaluations of item fit, we deleted three items: item 2 had $P < 0.001$ for the Muraki fit test and the difficulty level and content of the item seemed to be covered by several better fitting items; item 48 had $P < 0.001$, $RMSE = 0.63$ (on a 0–3 scale) and for 13% of the patients, the observed and expected responses differed more than one response category; item 51 had $P < 0.001$ and disordered threshold parameters.

We recalibrated the GPCM to the remaining 31 items. The parameter estimates and fit statistics are summarized in Table 4. Item 35 had $P < 0.001$, but since this was one of the few items relevant for patients with good PF, and the bias and RMSE were relatively small, we decided to keep the item. Otherwise, the fit of these 31 items appeared satisfactory.

The 31 items appeared to provide good content coverage of the PF aspects of interest and there was good variability in the difficulties of the items, although the item pool has relatively few items at the upper extreme. Figure 1 shows the total information for the 31 items. The plot shows that the items provide good information for a wide range of PF (PF scores around -2.5 to 0.5 , roughly corresponding to 10–80 if transformed to a 0–100 scale). The measurement precision is lower only for patients at the extremes, particularly for patients with good PF. For comparison, the information of the five QLQ-C30 PF items is also shown. As can be seen, these have markedly lower information than the total item pool across the whole PF continuum.

Table 4 Parameter estimates, number of effective categories, and fit statistics for the 31 items in the final IRT model

Item	No. of categories	Slope	Mean threshold	Item fit <i>P</i> -value	Bias	RMSE
Item 1	2	2.29	-1.99	0.213	0.000	0.18
Item 3	4	3.42	-1.33	0.003	0.004	0.39
Item 4	2	1.99	-0.99	0.177	0.002	0.34
Item 6	4	2.95	-1.21	0.054	0.004	0.45
Item 7	4	2.99	-2.01	0.001	0.002	0.22
Item 8	3	3.09	-1.91	0.130	0.001	0.19
Item 11	4	3.05	-0.87	0.286	0.005	0.50
Item 15	4	1.87	-0.68	0.049	0.006	0.68
Item 16	4	3.33	-0.39	0.005	0.008	0.54
Item 18	4	3.27	-2.07	0.002	0.001	0.21
Item 21	4	3.07	-1.28	0.007	0.004	0.43
Item 25	3	3.42	-1.83	0.007	0.001	0.20
Item 26	4	1.09	-1.69	0.254	0.003	0.60
Item 29	2	2.83	-1.93	0.386	0.000	0.15
Item 30	4	2.94	-0.46	0.012	0.008	0.56
Item 32	4	1.86	-1.15	0.010	0.004	0.59
Item 33	4	2.15	-1.36	0.015	0.004	0.49
Item 35	4	2.61	0.32	<0.001	0.001	0.57
Item 37	4	3.52	-1.54	0.005	0.003	0.33
Item 41	4	4.16	-0.40	0.061	0.008	0.47
Item 42	2	3.28	-2.24	0.140	0.000	0.10
Item 45	4	2.89	-0.14	0.046	0.010	0.59
Item 46	4	3.61	0.04	0.071	0.012	0.51
Item 47	4	3.22	-0.45	0.083	0.007	0.55
Item 49	4	2.31	-0.63	0.057	0.007	0.63
Item 50	2	4.45	-1.16	0.031	0.001	0.22
Item 52	3	4.33	-1.81	0.002	0.001	0.15
Item 53	4	3.58	-0.54	0.034	0.007	0.51
Item 54	2	2.24	-2.41	0.793	0.000	0.13
Item 55	2	2.32	-1.52	0.019	0.001	0.25
Item 56	4	3.58	-0.37	0.050	0.008	0.53

**Fig. 1** Test information function for the 31 items in the final model and for the five EORTC QLQ-C30 PF items, respectively

For the scale based on the 31 items, 0.3% responded “very much” to all items (floor effect) and 14.1% responded “not at all” to all items (ceiling effect), respectively. For the QLQ-C30 PF scale, the floor and ceiling effects were 0.7 and 31.2%, respectively.

Test for differential item functioning (DIF)

There were no indications of significant and relevant non-uniform DIF. Table 5 presents the findings of significant and potentially relevant uniform DIF. There was no significant DIF with regard to cancer site, cancer stage, current treatment, education, work, or cohabitation, and only item 4 showed DIF with regard to age. Nine items showed gender DIF. There was a general trend that women found

Table 5 Results of DIF analysis for the 31 items in the final IRT model

Item	Age		Gender		Country	
	β (CI) ^a	<i>P</i> -value	β (CI) ^b	<i>P</i> -value	β (CI) ^c	<i>P</i> -value
Item 1						
Item 3					1.83 (1.05; 2.61)	<0.0001
Item 4	1.33 (0.32;2.34)	0.0006				
Item 6					-1.60 (-2.39; -0.82)	0.0008
Item 7					-2.29 (-3.69; -0.89)	<0.0001
Item 8					-3.94 (-5.79; -2.08)	<0.0001
Item 11			-1.16 (-1.52; -0.80)	<0.0001	1.41 (0.75; 2.08)	<0.0001
Item 15			1.89 (1.56; 2.22)	<0.0001	-1.72 (-2.44; -1.00)	<0.0001
Item 16					-1.03 (-1.74; -0.32)	<0.0001
Item 18						
Item 21					1.29 (0.35; 2.22)	0.0002
Item 25						
Item 26					-1.70 (-2.33; -1.08)	<0.0001
Item 29						
Item 30			1.32 (1.01; 1.64)	<0.0001	0.83 (0.14; 1.52)	<0.0001
Item 32			1.11 (0.75; 1.48)	<0.0001	-2.01 (-2.76; -1.27)	<0.0001
Item 33						
Item 35					-1.07 (-1.68; -0.44)	<0.0001
Item 37					1.77 (0.99; 2.55)	<0.0001
Item 41			0.86 (0.53; 1.19)	<0.0001	-1.75 (-2.36; -1.14)	<0.0001
Item 42						
Item 45					2.00 (1.30; 2.70)	<0.0001
Item 46					1.07 (0.34; 1.79)	<0.0001
Item 47			-0.74 (-1.05; -0.42)	<0.0001	1.27 (0.65; 1.90)	<0.0001
Item 49			1.47 (1.14; 1.79)	<0.0001	-1.84 (-2.56; -1.12)	<0.0001
Item 50						
Item 52						
Item 53			0.73 (0.41; 1.06)	<0.0001	1.18 (0.44; 1.92)	<0.0001
Item 54					-2.98 (-4.96; -1.00)	0.0003
Item 55						
Item 56			-0.92 (-1.25; -0.59)	<0.0001	1.36 (0.61; 2.11)	0.0005

Regression coefficients (95% CI) and *P*-values for the significant findings of uniform DIF

There was no significant DIF with regard to cancer site, cancer stage, current treatment, education, work, or cohabitation

^a Largest regression coefficient for an age group when comparing with patients ≥ 70 years. A coefficient >0 indicates patients ≥ 70 years are more likely to report problems on the item

^b A coefficient >0 indicates women are more likely to report problems on the item

^c Largest regression coefficient for a country when comparing with United Kingdom. A coefficient >0 indicates UK patients are more likely to report problems on the item

lifting or carrying objects more demanding than did men, but regarded walking and running as relatively less demanding than men. Of the 31 items, 21 showed potential problems with country DIF. There was a tendency for patients from the United Kingdom to respond differently to items using units in the metric system. Otherwise, there were no obvious trends for the findings of country DIF.

To evaluate the impact of the DIF findings for the estimation of the PF score, we compared the PF scores obtained with the model ignoring DIF with the model taking DIF into account for item 8 (largest DIF), item 7 (large DIF in the same direction as item 8), and items 15 and 49 (two largest gender DIF). The comparisons are summarized in Table 6. The DIF in the individual items

Table 6 Evaluation of the impact of DIF findings on the estimation of the PF score

Item	DIF	PF scores based on all 31 items				PF scores based on DIF item(s) only			
		Median diff.	Range of diff.	Corr.	>Median SE(θ) (%)	Median diff.	Range of diff.	Corr.	> Median SE(θ) (%)
7	Country	0.00	−0.04; 0.02	1.00	0	−0.02	−0.10; 0.34	0.99	0
8	Country	0.00	−0.03; 0.01	1.00	0	−0.02	−0.14; 0.37	0.98	0
7 + 8	Country	0.00	−0.05; 0.02	1.00	0	−0.03	−0.13; 0.43	0.99	0
15	Gender	0.00	−0.02; 0.04	1.00	0	0.08	−0.28; 0.14	0.98	0
49	Gender	0.00	−0.01; 0.02	1.00	0	0.08	−0.23; 0.12	0.99	0

Comparison of scores based on the model ignoring DIF and models taking DIF into account

had almost no effect when using all 31 items to estimate the PF score; the possible DIF effect was “neutralised” by the other items. In a CAT, only a subset of items will typically be used to estimate PF. However, even in the extreme case, using the DIF item only to estimate PF, the DIF had only minor impact on the estimation of PF: all correlations between PF scores based on the two models were ≥ 0.98 , median differences were < 0.10 , and all differences were below the median standard error (SE) for the PF estimate. However, when taking the possible DIF for items 7 and 8 into account simultaneously (“7 + 8” in Table 6), some differences were approaching the median SE (up to $0.9 \times$ median SE).

Evaluation of discarded items and final IRT calibration

The evaluations of the items discarded in the previous steps indicated that adding any of these to the model again would result in significantly poorer model fit. Therefore, no items were reinstated. The 31 items included in the final model and their parameter estimates are shown in Table 4. Results of an eigenvalue analysis for the 31 items are shown in Table 3.

Discussion

We have reported on the selection and calibration of items for the EORTC PF item pool. The items of the original QLQ-C30 PF scale are somewhat heterogeneous capturing several important aspects of PF. We have previously reported on evidence of DIF for this scale [34]. Hence, we anticipated that it would be difficult, if not impossible, to find a perfectly unidimensional, DIF-free construct covering the same aspects as the QLQ-C30 PF scale. The aim was to generate an item pool covering the same aspects of PF as the QLQ-C30 scale, while maintaining a relatively simple and unidimensional structure. For this, we chose a pragmatic approach: accepting some deviations from “perfect fit” provided the deviations did

not seem to have substantial impact on the estimation of the PF score.

Using standard criteria for evaluating factor analytic results, our item pool had an acceptable fit to a unidimensional model. The fit statistics observed for our item pool are comparable to or better than those reported in similar studies on the development of CAT for measuring HRQOL domains [16, 17, 35].

Because of multiple testing, and because fit statistics are highly sensitive to sample size [8], we required clear indications of misfit to the IRT model ($P < 0.001$) before considering removing an item from the item pool. For example, taking a random subsample of 400 patients (approximately a third of our sample), all item fit statistics had $P \geq 0.02$ and all except four were > 0.05 , while in the full sample, 17 items had $P < 0.05$. Furthermore, the model was able to predict item responses quite well with very little systematic bias and only minor noise. We concluded that the fit of the 31 items in the final model was adequate, and are confident that possible deviations from perfect fit do not affect the estimation of PF.

About two-thirds of the items showed indications of large DIF. However, even the most pronounced DIF findings, clearly exceeding previously used cut-offs for large and relevant DIF [27, 28, 34], were found to have negligible impact on the estimation of PF. However, we did find that, if PF was estimated from two or more items with the same kind of DIF, it may result in relevant bias in the PF score. To avoid such problems, the CAT algorithm can be programmed to avoid posing only questions with similar DIF. Instead, one can counterbalance a DIF item with an item without DIF or with DIF in the opposite direction.

For a questionnaire such as the QLQ-C30, which includes a limited number of items per HRQOL domain, CAT has the potential for significantly improving measurement precision, even with relatively small item pools. The PF item pool includes 31 items, six times as many as the original scale. We found that the item pool provides good information for a wide range of PF scores and good coverage of the PF aspects assessed in the QLQ-C30.

However, for patients with good PF, the item pool has suboptimal measurement precision. That is, the item pool may lack items to distinguish precisely between patients with good and very good PF. Such items may be added at a later stage, but in many clinical studies, it may not be important to be able to distinguish precisely between patients with good and very good PF, since neither needs treatment nor intervention. Further, we would note that also at the extremes, the new item pool has markedly better precision than the original QLQ-C30 PF scale. The precision and efficiency of the PF CAT, and its acceptability by patients, will be further evaluated in future analyses. Because the item pool covers the same PF aspects and includes the QLQ-C30 PF items, scores based on the PF CAT can be compared with the large extant literature using the original QLQ-C30 scale.

We have restricted the current study to the development of CAT for the PF aspects measured by the QLQ-C30 and have retained the same item format as the QLQ-C30. This will ensure backward compatibility with the QLQ-C30, and for the many users familiar with QLQ-C30, the new instrument will be within a well-known measurement framework. However, the CAT instrument will not cover potentially relevant PF aspect not included in the QLQ-C30, and the restriction of the item format may limit the number of items that can be included in the item pool. The CAT instrument developed in the current study should be regarded as a first important step in the development of a new EORTC instrument, which may at later stages include other PF aspects and/or other response formats if this is deemed relevant.

Other CAT measures of PF are being developed. Most of these have focused on a specific patient population or type of impairment. This includes CAT for assessing PF in palliative care [36], in post-acute care rehabilitation [37], in children with cerebral palsy [38], and in patients with foot or ankle impairments [39], shoulder impairments [40], or lumbar spine impairments [41]. To our knowledge, our effort is the first to develop a CAT-based HRQOL measurement system for patients with cancer in general, in an international context.

The US-based Patient-Reported Outcomes Measurement Information System (PROMIS) project is developing CAT for the measurement of several HRQOL domains including PF for use across a wide range of chronic diseases [42, 43]. It will be of interest to compare the performance of the PROMIS measures to that of other CAT-based measures that have been developed for specific patient populations.

In conclusion, we believe that we have established a useful and suitable item pool for CAT measurement of PF that will clearly improve the EORTC measurement of PF. Together with a previous publication presenting the initial steps of this developmental process [6], the current paper

provides a complete picture of the EORTC approach to generating CAT versions of the HRQOL domains in the QLQ-C30. The methods used and experience gained in developing this CAT-based measure will serve as the basis for the development of item pools and CAT versions for the remaining QLQ-C30 domains.

Acknowledgments The study was funded by grants from the EORTC Quality of Life Group. National Taiwan University, grant National Science Council, Taiwan, No. 95-2314-B-002-266-MY2, 97-2314-B-002-020-MY3.

References

1. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The European Organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85(5), 365–376.
2. Fayers, P., & Bottomley, A. (2002). Quality of life research within the EORTC—the EORTC QLQ-C30. European Organisation for research and treatment of cancer. *European Journal of Cancer*, 38(Suppl 4), S125–S133.
3. Garratt, A., Schmidt, L., Mackintosh, A., & Fitzpatrick, R. (2002). Quality of life measurement: Bibliographic study of patient assessed health outcome measures. *British Medical Journal*, 324(7351), 1417–1419.
4. Fayers, P. M., Aaronson, N. K., Bjordal, K., Groenvold, M., Curran, D., & Bottomley, A. (2001). *The EORTC QLQ-C30 scoring manual*. Brussels: European Organisation for Research and Treatment of Cancer.
5. Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum.
6. Petersen, M. A., Groenvold, M., Aaronson, N. K., Chie, W.-C., Conroy, T., Costantini, A., et al. (2010). Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions—General approach and initial results for physical functioning. *European Journal of Cancer*, 46, 1352–1358.
7. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.
8. van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Berlin: Springer.
9. Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement*, 1(2), 114–142.
10. Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. *Quality of Life Research*, 16(Suppl 1), 187–194.
11. Muraki, E., & Bock, R. D. (1996). *PARSCALE—IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago: Scientific Software International, Inc.
12. Muthen, L. K., & Muthen, B. O. (2002). *Mplus user's guide*. Los Angeles, CA: Muthen & Muthen.
13. Cattell, R. B. (1966). Scree test for number of factors. *Multivariate Behavioral Research*, 1(2), 245–276.
14. Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
15. Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

16. Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the headache impact test (HIT). *Quality of Life Research*, *12*(8), 913–933.
17. Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, *14*(10), 2277–2291.
18. Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*(1), 63–79.
19. Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–168). Berlin: Springer.
20. Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. V. D. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). Berlin: Springer.
21. Samejima, F. (1997). Graded response model. In W. V. D. Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100).
22. Maydeuolivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, *18*(3), 245–256.
23. Hanson, B. A. (2009). IRT command language (ICL) program <http://www.b-a-h.com/software/irt/icl>. Accessed June 2009.
24. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, *39*, 1–22.
25. Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.
26. French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315–332.
27. Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N. K., Conroy, T., Cull, A., et al. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research*, *12*(4), 373–385.
28. Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology*, *51*(11), 1189–1202.
29. Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *Alberta Journal of Educational Research*, *45*(4), 353–376.
30. Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.
31. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2009). The practical impact of differential item functioning analyses in a health-related quality of life instrument. *Quality of Life Research*, *18*(8), 1125–1130.
32. Hart, D. L., Deutscher, D., Crane, P. K., & Wang, Y. C. (2009). Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. *Quality of Life Research*, *18*(8), 1067–1083.
33. SAS Institute Inc. (2004). *SAS/STAT[®] 9.1 user's guide*. Cary, NC: SAS Institute Inc.
34. Scott, N. W., Fayers, P. M., Bottomley, A., Aaronson, N. K., de Graeff, A., Groenvold, M., et al. (2006). Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research*, *15*(6), 1103–1115.
35. Revicki D. A., Chen, W. H., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., et al. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain*.
36. Helbostad, J. L., Holen, J. C., Jordhoy, M. S., Ringdal, G. I., Oldervoll, L., & Kaasa, S. (2009). A first step in the development of an international self-report instrument for physical functioning in palliative cancer care: A systematic literature review and an expert opinion evaluation study. *Journal of Pain and Symptom Management*, *37*(2), 196–205.
37. Haley, S. M., Ni, P., Hambleton, R. K., Slavin, M. D., & Jette, A. M. (2006). Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. *Journal of Clinical Epidemiology*, *59*(11), 1174–1182.
38. Haley, S. M., Fragala-Pinkham, M. A., Dumas, H. M., Ni, P., Gorton, G. E., Watson, K., et al. (2009). Evaluation of an item bank for a computerized adaptive test of activity in children with cerebral palsy. *Physical Therapy*, *89*(6), 589–600.
39. Hart, D. L., Wang, Y. C., Stratford, P. W., & Mioduski, J. E. (2008). Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function. *Quality of Life Research*, *17*(8), 1081–1091.
40. Hart, D. L., Cook, K. F., Mioduski, J. E., Teal, C. R., & Crane, P. K. (2006). Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology*, *59*(3), 290–298.
41. Hart, D. L., Mioduski, J. E., Werneke, M. W., & Stratford, P. W. (2006). Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology*, *59*(9), 947–956.
42. Fries, J. F., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *Journal of Rheumatology*, *36*(9), 2061–2066.
43. Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supports the expected advantages of the patient-reported outcomes measurement information system (PROMIS). *Journal of Clinical Epidemiology*, *61*(1), 17–33.