The Dutch–Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain

Crins, M.H.P.; Terwee, C.B.; Klausch, T.; Smits, N.; de Vet, H.C.W. ; Westhovens, R.; Cella, D.; Cook, K.F.; Revicki, D.A.; Leeuwen, J. van; Boers, M.; Dekker, J.; Roorda, L.D.

DOI
10.1016/j.jclinepi.2017.03.011

Publication date
2017

Document Version
Final published version

Published in
Journal of Clinical Epidemiology

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 87 (2017) 47–58

Journal of Clinical Epidemiology

The Dutch–Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain

Martine H.P. Crins[a,*], Caroline B. Terwee[b], Thomas Klausch[b], Niels Smits[c], Henrica C.W. de Vet[b], Rene Westhovens[d,e], David Cella[f], Karon F. Cook[f], Dennis A. Revicki[g], Jaap van Leeuwen[h], Maarten Boers[b,i], Joost Dekker[j,k], Leo D. Roorda[a]

[a]*Amsterdam Rehabilitation Research Center|Reade, Doctor Jan van Breemenstraat 2, Amsterdam 1056 AB, The Netherlands*
[b]*Department of Epidemiology and Biostatistics, The EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, Amsterdam 1081, The Netherlands*
[c]*Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, Amsterdam 1018 WS, The Netherlands*
[d]*Department of Development and Regeneration, Skeletal Biology and Engineering Research Center, KU Leuven, Herestraat 49, Leuven 3000, Belgium*
[e]*Rheumatology, University Hospitals, KU Leuven, Herestraat 49, Leuven 3000, Belgium*
[f]*Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 633 N. Saint Clair Street, 19th Floor, Chicago, IL 60611, USA*
[g]*Outcomes Research, Evidera, 7101 Wisconsin Ave., Suite 1400, Bethesda, MD 20814, USA*
[h]*Leones Group BV, Middenweg 78, Dirkshorn 1746 EB, The Netherlands*
[i]*Amsterdam Rheumatology and Immunology Center, VU University Medical Center, De Boelelaan 1117, Amsterdam 1081 HV, The Netherlands*
[j]*Department of Rehabilitation Medicine, VU University Medical Center, De Boelelaan 1117, Amsterdam 1081 HV, The Netherlands*
[k]*Department of Psychiatry, VU University Medical Center, De Boelelaan 1117, Amsterdam 1081 HV, The Netherlands*

Accepted 17 March 2017; Published online 28 March 2017

Abstract

Objective: The objective of this study was to assess the psychometric properties of the Dutch–Flemish Patient-Reported Outcomes Measurement Information System (PROMIS) Physical Function item bank in Dutch patients with chronic pain.

Study Design and Setting: A bank of 121 items was administered to 1,247 Dutch patients with chronic pain. Unidimensionality was assessed by fitting a one-factor confirmatory factor analysis and evaluating resulting fit statistics. Items were calibrated with the graded response model and its fit was evaluated. Cross-cultural validity was assessed by testing items for differential item functioning (DIF) based on language (Dutch vs. English). Construct validity was evaluated by calculation correlations between scores on the Dutch–Flemish PROMIS Physical Function measure and scores on generic and disease-specific measures.

Results: Results supported the Dutch–Flemish PROMIS Physical Function item bank's unidimensionality (Comparative Fit Index = 0.976, Tucker Lewis Index = 0.976) and model fit. Item thresholds targeted a wide range of physical function construct (threshold-parameters range: −4.2 to 5.6). Cross-cultural validity was good as four items only showed DIF for language and their impact on item scores was minimal. Physical Function scores were strongly associated with scores on all other measures (all correlations ≤ -0.60 as expected).

Conclusion: The Dutch–Flemish PROMIS Physical Function item bank exhibited good psychometric properties. Development of a computer adaptive test based on the large bank is warranted.

Keywords: Physical functioning; Chronic pain; PROMIS; Dutch–Flemish PROMIS; Item response theory; Psychometrics

1. Introduction

Physical function refers to the ability to perform activities of daily living and instrumental activities of daily living [1]. Limitations in physical function are a big concern for elderly and patients with chronic diseases such as musculoskeletal disease and chronic pain [2–4]. Because physical function is prerequisite for independent

Funding: The Dutch–Flemish translation of the PROMIS item banks was supported by a grant from the Dutch Arthritis Association (BP 10-1-261). Funding for the current calibration study was provided by a grant from the Dutch Scientific College of Physical Therapy.

* Corresponding author. Tel.: +31-20-589-6544; fax: +31-20-489-6704.

E-mail address: m.crins@reade.nl (M.H.P. Crins).

http://dx.doi.org/10.1016/j.jclinepi.2017.03.011
0895-4356/

What is new?

- Findings from this study in patients with chronic pain contribute to the evidence for the good psychometric properties of the Patient-Reported Outcomes Measurement Information System (PROMIS) Physical Function item bank.
- If applied as Short Form or computerized adaptive testing (CAT), the item bank has the potential to improve the measurement of physical function in a user-friendly and efficient way.
- Results from this study expand the evidence for the validity of the Dutch−Flemish PROMIS Physical Function item bank across populations.

living, it is commonly measured in clinical care and often is a core outcome of treatment [3]. A large number of patient-reported outcome measures (PROMs) are available to measure physical function in patient populations [5−9]. However, these PROMs can be burdensome for patients because of their length and because items are not targeted to respondents' individual levels of physical function. Existing measures vary in measurement quality and precision and may have limited measurement range (i.e., ceiling and floor effects). Furthermore, scores of different physical function measures are not comparable across different PROMs.

The National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS) initiative developed a new state-of-the-art generic assessment tool to measure patient-reported health across different populations. To optimize content validity, items adapted from existing PROMs and new items were combined into item banks [10−14]. An item bank is a set of items (questions) measuring a common construct such as physical function [15]. Responses to items in an item bank were calibrated with item response theory (IRT), which orders items along a measurement continuum, based on item difficulty (e.g., ''are you able to run a mile'' is a more difficult item than ''are you able to move across the room'') and discrimination [16]. Once calibrated to an IRT model, the item bank can be used to tailor measurement to individual persons through computerized adaptive testing (CAT) [11]. A CAT is a computer-administered measure in which successive items are selected by a computer algorithm based on responses to previous items [11,15,17]. After each item, the person's score and the associated standard error are estimated, and when a predefined precision (e.g., standard error <0.3 on the theta metric; <3 on the T-score metric) is achieved, the computer stops administering items and estimates the final score. Typically, this ''stopping criterion'' can be reached after administering 3 to 7 items. It can also be programmed that the computer stops administering items after a certain number of items are administered, also called fixed-length CAT. Because the administration of items is tailored to individuals, persons only respond to a minimal number of highly informative and relevant items. With CATs, higher measurement precision (less measurement error) can be achieved with less response burden.

PROMIS item banks and CATs have the potential to be implemented worldwide. The PROMIS Physical Function item bank has shown to have stronger content validity, better responsiveness, and other desirable psychometric properties compared with traditional physical function PROMs such as the SF-36 Health Survey Physical Functioning scale (SF-36 PF) and the Health Assessment Questionnaire−Disability index (HAQ-DI) [12−14,18−20]. Furthermore, PROMIS scores are expressed on a standardized T-score metric (T-score 50 represents the average score of the general US population, with a standard deviation of 10) that facilitates interpretability of scores [21,22].

The Dutch−Flemish PROMIS Group translated 17 adult PROMIS item banks (including the PROMIS Physical Function item bank) and nine pediatric PROMIS item banks into Dutch−Flemish [23,24]. The Dutch−Flemish PROMIS Physical Function item bank has been administered and tested in patients with rheumatoid arthritis, which is a relatively homogeneous patient group [25]. In line with the PROMIS goals to calibrate a translated item bank in multiple validation studies and in patients with multiple conditions, we conducted a second validation study with the Dutch−Flemish PROMIS Physical Function item bank in a more heterogeneous sample, to evaluate the generalizability of its properties across patient populations.

The objective of present study was to examine unidimensionality and calibrate the item parameters of the Dutch−Flemish PROMIS Physical Function item bank in Dutch chronic pain patients. Furthermore, the objective was to evaluate the cross-cultural validity of the Dutch−Flemish compared with the US PROMIS Physical Function item bank, and its reliability and construct validity. The ultimate aim was to obtain a valid, reliable, user-friendly, and efficient measurement tool for assessing physical function, available to care providers and researchers in both the Netherlands and Flanders (the Dutch-speaking part of Belgium).

2. Methods

2.1. Study participants

For this study, 2,808 patients from the Amsterdam Pain (AMS-PAIN) cohort were invited to participate. The AMS-PAIN cohort comprises chronic pain patients who have been registered since September 2010 in Reade, an outpatient secondary care center for rheumatology and rehabilitation in the Netherlands. To be eligible, patients had to have at least one chronic pain condition of the

musculoskeletal system for at least 3 months before participating in the study, had to be 21 years or older, and had to provide informed consent.

To evaluate the cross-cultural validity (or equivalence) of the Dutch−Flemish compared with the US PROMIS Physical Function item bank, data from a US PROMIS wave 1 subsample (the full bank samples) were used [12,14]. In this US PROMIS wave 1 subsample, the Physical Function item bank was divided into two item blocks (containing 95 and 26 of the 121 items, respectively), which were completed by two independent samples of 942 and 995 participants, respectively (total 1937 US participants). Of the 1,937 respondents, 237 participants were excluded because of inappropriate response time, which resulted in 1,700 US participants for cross-cultural validity analysis.

2.2. Procedures

The study was approved by the local institutional review board (Slotervaart hospital and Reade). Patients from the AMS-PAIN cohort were invited by e-mail or letter, to complete a Web-based (digital) or paper-and-pencil (paper) questionnaire that included, among other measures, the full Dutch−Flemish PROMIS Physical Function item bank.

2.3. Measures

The questionnaire included questions addressing demographic (i.e., age, gender, country of birth, educational level) and clinical characteristics (i.e., duration of pain, type chronic pain condition). These questions enabled description of the Dutch−Flemish sample and also comparison to the US sample.

The questionnaire also included all 121 items of the Dutch−Flemish PROMIS Physical Function item bank [23,26]. There are three items (with the word ''walking'' in it) that have a slightly different translation in the Dutch and Flemish language. In the present study, the Dutch version of these items was used. The items cover a wide range of activities, from self-care (activities of daily living) to more complex activities that require a combination of skills [1,12−14,27]. There is no time frame for the items, but current status is inferred. The item bank includes items about functioning of the axial regions (neck and back), the upper and lower extremities, and ability to carry out instrumental activities of daily living (i.e., housework, shopping) [1,13,14,27]. There are three different five-point Likert response scales: (1) unable to do/with much difficulty/with some difficulty/with a little difficulty/without any difficulty; (2) cannot do/quite a lot/somewhat/very little/not at all; and (3) cannot do because of health/a lot of difficulty/some difficulty/a little bit of difficulty/no difficulty at all. Higher scores indicate better function. For the present study, the item bank was split-up into three parts to prevent burden on respondents by answering 121 items in succession about the same topic.

Finally, the questionnaire included two generic and four disease-specific instruments to assess construct validity. The generic instruments were applied in all patients. They consisted of the Dutch−Flemish PROMIS Pain Interference item bank, and the PROMIS pain intensity item (Global07) from the Dutch−Flemish PROMIS Global Health item bank. The disease-specific instruments were applied in patients with chronic neck, shoulder, back or widespread pain, respectively. They consisted of the ''Neck Disability Index'' (NDI), the ''Disabilities of the Arm, Shoulder and Hand'' (DASH) questionnaire, the ''Roland Morris Disability Questionnaire'' (RMDQ), and the ''Fibromyalgia Impact Questionnaire'' (FIQ). The Dutch−Flemish PROMIS Pain Interference item bank consists of 40 items and assesses the consequences of pain on relevant aspects of persons' lives and includes the impact of pain on social, cognitive, emotional, physical, and recreational activities as well as sleep and enjoyment in life [1,28,29]. The scores are expressed in T-scores. The Dutch−Flemish PROMIS Pain Interference item bank was evaluated in Dutch chronic pain patients and showed good psychometric properties [29]. The PROMIS pain intensity item (Global07) from the Dutch−Flemish PROMIS Global Health item bank consist of an 11-point numeric rating scale (NRS) with anchors 0 = ''no pain'' and 10 = ''worst pain imaginable'' [30,31]. The NDI consists of 10 items measuring self-reported pain intensity and the influence of neck pain on daily activities, with a total score ranging from 0 to 50 [32,33]. Evidence has accumulated for the test−retest reliability and validity of the NDI within Dutch patients with chronic neck pain [33−36]. The DASH questionnaire consists of 30 items measuring disabilities of the upper extremities, with a total score ranging from 0 to 100 [37,38]. DASH scores have demonstrated good reliability and good validity (i.e., strong correlations with instruments measuring related constructs) in Dutch patients with a variety of unilateral disorders of the upper limb [38−41]. The RMDQ consists of 24 items measuring disabilities as a result of chronic back pain, with a total score ranging from 0 to 24 [42,43]. RMDQ scores have demonstrated good reliability and validity within Dutch patients with chronic low back pain [42,44−46]. The FIQ consists of 20 items measuring physical disabilities as a result of fibromyalgia, with a total score ranging from 0 to 100 [47,48]. FIQ scores have demonstrated moderate to good reliability and validity among Dutch patients with fibromyalgia [48,49]. For each instrument, higher scores indicate more interference, intensity, disability, or impact.

The total survey comprised of demographic and clinical questions, followed by the total Pain Behavior item bank (39 items, but analyses are not included in current study), part 1 of the PF item bank (50 items), the total Pain Interference item bank (40 items), part 2 of the PF item bank (46 items), the Global Health item bank (10 items), part 3 of the PROMIS PF item bank (25 items), and last the four disease-specific instrument questionnaires NDI, DASH, RMDQ, and FIQ.

2.4. Statistical analysis

2.4.1. Study participants

Demographic and clinical characteristics were described by descriptive statistics. Differences between the Dutch AMS-PAIN sample and the US wave 1 sample were evaluated by independent sample *t*-tests for continuous variables and chi-square tests for categorical variables.

2.4.2. Calibration of the Dutch−Flemish PROMIS Physical Function item bank

Psychometric analyses were conducted following the PROMIS analysis plan and were similar to those used in the calibration of the Dutch−Flemish PROMIS Pain Behavior item bank and the Dutch−Flemish PROMIS Pain Interference item bank [29,50,51].

First, unidimensionality was examined by confirmatory factor analyses (CFAs) on the polychoric correlation matrix. All items were hypothesized to load on a single factor. The analysis was performed with the R-package (version 3.0.1) Lavaan (version 0.5−16), and model fit was evaluated based on the Comparative Fit Index (CFI), Tucker Lewis Index (TLI), and Root Means Square Error of Approximation (RMSEA) [52,53]. The criteria for unidimensionality include CFI $>$ 0.95, TLI $>$ 0.95, and RMSEA $<$ 0.06 [50]. Furthermore, unidimensionality was considered sufficient when with exploratory factor analysis (EFA) the first factor accounts for at least 20% of the variability and when the ratio of the variance explained by the first to the second factor is greater than 4 [50,54].

Second, local independence was evaluated. After controlling for the dominant factor, there should be no significant covariance among item responses. This was evaluated by examining the residual correlation matrix resulting from a single factor CFA of all item responses. Residual correlations greater than 0.2 were considered as indicators of possible local dependence [50]. The impact of local dependency on IRT parameter estimates was evaluated, by removing one item of a locally dependent pair and examining changes in the IRT parameters of the remaining items [11].

Third, monotonicity was assessed. The probability of endorsing a higher item response category should increase (or at least not decrease) with increasing levels of the underlying construct. Monotonicity of the Dutch−Flemish PROMIS Physical Function items was evaluated by fitting a nonparametric IRT model, with Mokken scaling in the R-package Mokken [55−58]. This model yields nonparametric IRT response curve estimates, shows the probabilities of endorsing response categories, and can be visually inspected to evaluate monotonicity. Fit of the monotone homogeneity model was evaluated by calculating the scalability coefficient H. Scale criteria are met if (1) the scalability coefficients for all item pairs (H_{ij}) are positive, (2) the scalability coefficients for the items in relation to the scale at issue (H_i) are at least 0.30, and (3) the scalability coefficient for the scale (H) is at least 0.30. Higher values for H_i and H indicate a better scale. A rule of thumb is that a scale is considered to be strong when H is $\geq$0.50.

After evaluation of these three assumptions, a graded response model (GRM) was fit to the data to calibrate the item parameters with the R-package MIRT [59,60]. The GRM models two item parameters, the item slope and the item threshold [50]. The item slope parameter estimates the discriminative ability of the items, with higher slope values indicating better ability to discriminate between adjoining values on the construct. Item threshold parameters estimate item difficulty and locate the items along the measured trait. To assess the fit of the GRM and the degree to which possible misfit affects the IRT model, S-X^2 statistic was used [61]. This statistic compares the observed and expected response frequencies under the estimated IRT model and quantifies the differences between the observed and expected response frequencies. The criterion for poor fit of an item is an S-X^2 *P*-value of less than 0.001 [50,61].

2.4.3. Differential item functioning

Differential item functioning (DIF) analyses examine whether people from different groups (e.g., age or gender) with the same level of trait (in this study, the same level of physical function) have different probabilities of giving a certain response to an item [16,50,62]. There are two kinds of DIF: uniform and nonuniform [16,50,62]. Uniform DIF exists when the DIF is consistent, with the same magnitude of DIF across the entire range of the trait. Nonuniform DIF exists when the magnitude or direction of DIF differs across the trait. DIF was evaluated by ordinal logistic regression models in the R-package Lordif (version 0.3−3), in which a McFadden's pseudo R^2 change of 2% was used as the critical value to flag for possible DIF [50,63−65]. Within the Dutch AMS-PAIN sample, DIF was evaluated for age (median split: under 50 years vs. 50 years and over), gender (male vs. female), and administration mode (digital vs. paper).

2.4.4. Cross-cultural validity

Cross-cultural validity of the Dutch−Flemish PROMIS Physical Function item bank vs. the US PROMIS Physical Function item bank was assessed to examine if Dutch and US participants with the same level of trait (physical function) have different probabilities of given responses to an item. DIF for language (Dutch vs. English) was analyzed as described under Section 2.4.3. When items were flagged as potential DIF for language items, the impact of DIF was examined by plotting item characteristic curves (ICCs) and test characteristic curves (TCCs). The TCC plots showed the total item scores for all 121 PROMIS Physical Function items (ignoring DIF) and the scores for only the items having DIF [64,65].

2.4.5. Reliability

Reliability within IRT is conceptualized as "information," in which the fact that measurement precision can differ across levels of the measured trait (θ = theta) is taken into account. The relationship between information and standard error (SE) is defined by the formula: $SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$, where SE is the standard error of estimated θ, I is information, and θ is the estimated trait level (ranging from low levels to high levels of physical function) [28,66]. The formula indicates that increased scale information is related to smaller SEs and, therefore, greater measurement precision. With the calculated SEs, plots were overlaid showing SEs, as a parameter of reliability, across the score range of the total Dutch–Flemish PROMIS Physical Function item bank (121 items), the 4-, 6-, 8-, 10-, and 20-item Short Forxms (similar to the US short forms), and the 4-, 6-, 8-, 10-, and 20-item simulated fixed-length CATs (always applying 4, 6, 8, 10, and 20 items, respectively, no other stopping rules). The simulated fixed-length CATs were conducted with use of the R-package catR (version 3.4) [67]. With these plots, it is also shown how the total item bank vs. short forms vs. fixed-length CATs perform compared with each other. The IRT theta scores of the Dutch AMS-PAIN sample were, among others for reasons of international comparability, transformed into T-scores anchored on the US item parameters of the PROMIS Physical Function item bank. For the same reason, the plots of the reliability of the total item bank, the T-scores of the short forms, and the simulated fixed-length CATs were also based on US item parameters. The US item parameters can be found in the US PROMIS Physical Function "cue sheet"; the spreadsheet listing the items, response options, and item parameters of the US PROMIS Physical Function item bank. This cue sheet can be requested on the US PROMIS web site www.healthmeasures.net, at which up-to-date PROMIS tools and assistance are provided. The T-score 50 represents the average score of the general US population, with a standard deviation of 10.

2.4.6. Construct validity

T-scores of the Dutch–Flemish PROMIS Physical Function item bank were correlated (Pearson correlations) to the scores on the two generic instruments (the PROMIS Pain Interference item bank and PROMIS Pain Intensity item) and the four disease-specific instruments (the NDI, DASH, RMDQ, and FIQ). For assessing convergent validity, we hypothesized that the Physical Function item bank scores would have the strongest negative correlation ($r < -0.70$) with the Pain Interference item bank T-score because in patients with chronic pain, we hypothesized little distinction between the concepts physical function and pain interference. Furthermore, we expected strong negative correlations ($r < -0.50$) of the Physical Function item bank score with the four disease-specific disability instruments. We expect slightly lower correlations with the disability instruments than with the Pain Interference item bank because the disability instruments measure the concept of disability more broadly (e.g., also including items on pain and other symptoms). For assessing discriminant validity, we expected a moderate negative correlation (r between −0.30 and −0.50) between the Physical Function item bank and the Pain Intensity item score because pain intensity is a different construct than physical function.

3. Results

3.1. Study participants

Of the 2,808 invited patients, 1,286 completed the questionnaire (response rate 46%, 56% digital and 44% paper). No differences were found between responders and nonresponders with respect to age, gender, country of birth, or education level. Among the 1,286 respondents, 29 patients were excluded because they did not give informed consent and 10 patients were excluded because responses to all items of the Physical Function item bank were missing, resulting in 1,247 patients participating in the study. Most statistical analyses were performed based on data from the 921 respondents who had complete data on all items of the Physical Function item bank. However, because the GRM analyses can accommodate incomplete data, data of all 1,247 patients were used for the IRT calibration.

The demographic characteristics of the Dutch AMS-PAIN sample and the US wave 1 sample are summarized in Table 1. Of the AMS-PAIN patients, 78% were female, the average age (standard deviation [SD]) was 48 years (12) with a range from 21 to 85, 81% indicated that the duration of their pain was more than 2 years, and the average pain intensity on an NRS (SD) was 6.7 (2). The results of the t-tests and chi-square tests showed that the Dutch AMS-PAIN patients were slightly younger, a larger proportion was female, had higher pain intensity scores, and were less educated than the US wave 1 sample.

3.2. Calibration of the Dutch–Flemish PROMIS Physical Function item bank

The CFA results indicated good fit to a unidimensional model. The CFI was 0.976 and the TLI was 0.976, which are larger than the criterion of >0.95 [50]. The RMSEA was 0.122, which is larger than the criterion of <0.06. Notably, this criterion was not met in the US PROMIS Physical Function data either, nor many of the other PROMIS item banks. Reise et al. have found the RMSEA statistic to be problematic for assessing unidimensionality of health concepts [68]. To further investigate the dimensionality of the response data, we conducted an EFA. The first factor in EFA accounted for 56.6% of the variance, and the second factor accounted for 8.2% of the variance; hence, the ratio of the variance explained by the first to the second factor is 6.9, which is higher than the

Table 1. Demographic and clinical characteristics of the Dutch AMS-PAIN sample (n = 1,247) and the US wave 1 sample (n = 1,700)

	Dutch chronic pain sample	US wave 1 sample
Age mean (SD), range	48 (13), 21–85	51 (18), 18–88*
Gender, *n* (%)		
Male	278 (22)	824 (49)*
Female	969 (78)	876 (51)*
Country of birth, *n* (%)		
Netherlands	633 (51)	NA
Turkey	90 (7)	
Morocco	102 (8)	
Surinam	74 (6)	
Other	348 (28)	
Social status,[a] *n* (%)		
Single	432 (35)	NA
Married or living together	667 (54)	
Living apart together	65 (5)	
Living with parents	26 (2)	
Other	66 (5)	
Educational level, *n* (%)		
Less than high school degree	190 (18)	31 (2)*
High school degree	150 (14)	298 (17)
Some college	420 (41)	656 (39)
College degree	47 (5)	427 (25)*
Advanced degree	221 (22)	285 (17)*
Employment status,[a] *n* (%)		
Full time	203 (16)	NA
Part time	315 (25)	
Student	45 (4)	
Unpaid, volunteer, household	181 (15)	
Retired	103 (8)	
Unemployed	226 (18)	
Other	249 (20)	
Social benefits,[a] *n* (%)		
Sick listed	290 (23)	NA
Disability benefit	281 (23)	
Unemployment benefit	99 (8)	
Other	159 (13)	
No social benefit	432 (35)	
Duration of pain, *n* (%)		
3–6 months	18 (2)	NA
6–12 months	51 (4)	
1–2 years	163 (13)	
2–5 years	356 (29)	
>5 years	648 (52)	
Type of chronic pain condition,[a] *n* (%)		
Migraine and/or other "daily" headache	444 (36)	NA
Rheumatoid arthritis	155 (12)	
Osteoarthritis	424 (34)	
Pain related to cancer	20 (2)	
Lower back pain	878 (70)	
Neck or shoulder pain	872 (70)	
Fibromyalgia	422 (34)	
Chronic widespread pain	588 (47)	
Other neuropathic pain (nerve damage)	260 (21)	
Other	585 (47)	
No chronic pain condition	6 (0.5)	
T-score of the PROMIS Physical Function item bank		
Mean (SD)	35.7 (7.4)	50.8 (9.3)*
Range	10.2–73.5	12.4–71.3
Generic and disease-specific instruments mean (SD)		
PROMIS Pain Interference (n = 1,085)	64.1 (6.8)	—
PROMIS Global Health Pain intensity (n = 1,167)	6.7 (2)	2.5 (2)*
NDI (n = 448)	25 (9)	—

(*Continued*)

Table 1. Continued

	Dutch chronic pain sample	US wave 1 sample
DASH (n = 450)	47 (20)	—
RMDQ (n = 743)	13 (6)	—
FIQ (n = 337)	60 (18)	—

Abbreviations: AMS-PAIN, Amsterdam Pain; SD, standard deviation; NDI, Neck Disability Index (0–50); DASH, Disabilities of the Arm, Shoulder and Hand (0–100); NA, not applicable; RMDQ, Roland Morris Disability Questionnaire (0–24); FIQ, Fibromyalgia Impact Questionnaire (0–100).

PROMIS Pain Interference (T-score); PROMIS Global Health Pain Intensity (0–10). Higher scores indicate more interference, intensity, disability, or impact.

*$P < 0.001$.

[a] Multiple answers were allowed.

recommended minimum of 4 [50]. Based on these results, it was concluded that the responses to the Dutch–Flemish PROMIS Physical Function items were sufficiently unidimensional for calibration using IRT.

The residual correlation matrix showed a small number of deviances from local dependence among the items. Out of the 7,260 items pairs, 445 (6%) were flagged for local dependence. The item pairs with the greatest dependency were PFC33 (“Are you able to run 10 miles (16 km)?”)—PFC7 (“Are you able to run five miles (8 km)?”) with a residual correlation of 0.57, and PFA39 (“Are you able to run at a fast pace for two miles (3 km)?”)—PFC7 (“Are you able to run five miles (8 km)?”) with a residual correlation of 0.43. These items were removed sequentially and the items were recalibrated to evaluate the impact of local dependency on item parameter estimates. The mean differences in item thresholds estimates after removal of an item was −0.03, and the maximum difference was −0.08. The mean difference in slope parameter estimates was 0.014, and the maximum difference was 0.016. These results suggest minimal impact of local dependence.

Evaluation of monotonicity showed that the Dutch–Flemish PROMIS Physical Function item bank had only six significant violations (PFA10, PFA34, PFA38, PFB5, PFB10, PFC32). The scalability coefficient H_i of the items was ≥0.42 and higher than the lower bound of 0.30 for all items (details can be found in the online only Appendix 1 at www.jclinepi.com). The Mokken scalability coefficient H of the full item bank was 0.56, suggesting strong scalability. Based on these results, it was concluded that the Dutch–Flemish PROMIS Physical Function items sufficiently met the assumption of monotonicity.

The item slope parameters ranged from 0.8 to 3.1, with mean of 2.1. The item with lowest discrimination parameter was PFC33 (“Are you able to run ten miles (16 km)?”), and the item with the highest discrimination parameter was PFA16 (“Are you able to dress yourself, including tying shoelaces and buttoning your clothes?”). The item threshold parameters ranged from −4.2 to 5.6. There were five items with sparse item categories. In three items (PFA43, PFA50, and PFB16), there were few responses in categories 1–3, and in two items (PFC7 and PFC33), there were few responses in categories 3–5. The probability values for the S-X^2 statistics ranged from 0.0001 to 0.977. Based on the S-X^2 P-value of less than 0.001, only 1 of 121 items (PFB1) was found to misfit the GRM. From all calibration results, it was concluded not to remove items from the item bank. Details of the IRT item parameters can be found in the online only Appendix 1 at www.jclinepi.com.

3.3. Differential item functioning

None of the Dutch–Flemish PROMIS Physical Function items were flagged for DIF for gender or administration mode. One item (PFC33) was flagged for uniform DIF for age. The TCC (not shown) demonstrated negligible influence of the DIF for age item on the physical function total score, indicating negligible impact of DIF by age.

3.4. Cross-cultural validity

Four items showed some DIF for language (R^2 change values can be found in the online only Appendix 1 at www.jclinepi.com). All four were flagged for uniform DIF. For two items (PFB5 and PFC32), the Dutch patients were more likely to endorse higher response categories compared with the US participants, indicating that the items were easier for Dutch patients. For the other two uniform DIF items (PFB1 and PFB13), the Dutch patients were more likely to endorse lower response categories compared with the US participants.

The overall impact of DIF for language on the TCC is shown in Fig. 1, and the ICCs are not shown. The left graph shows the TCC for all 121 Physical Function items (ignoring DIF), and the right graph shows the TCC for the four items having DIF. These curves show that the physical function total score is only slightly different for Dutch patients than for US participants, indicating minimal impact of DIF by language.

3.5. Reliability

Because of the minimal impact of DIF by language and for the purpose of international comparability, the Dutch–Flemish T-scores were anchored on the item parameters of the US PROMIS Physical Function item bank. As shown in Table 1, the resulting mean T-score for the overall Dutch-Flemish

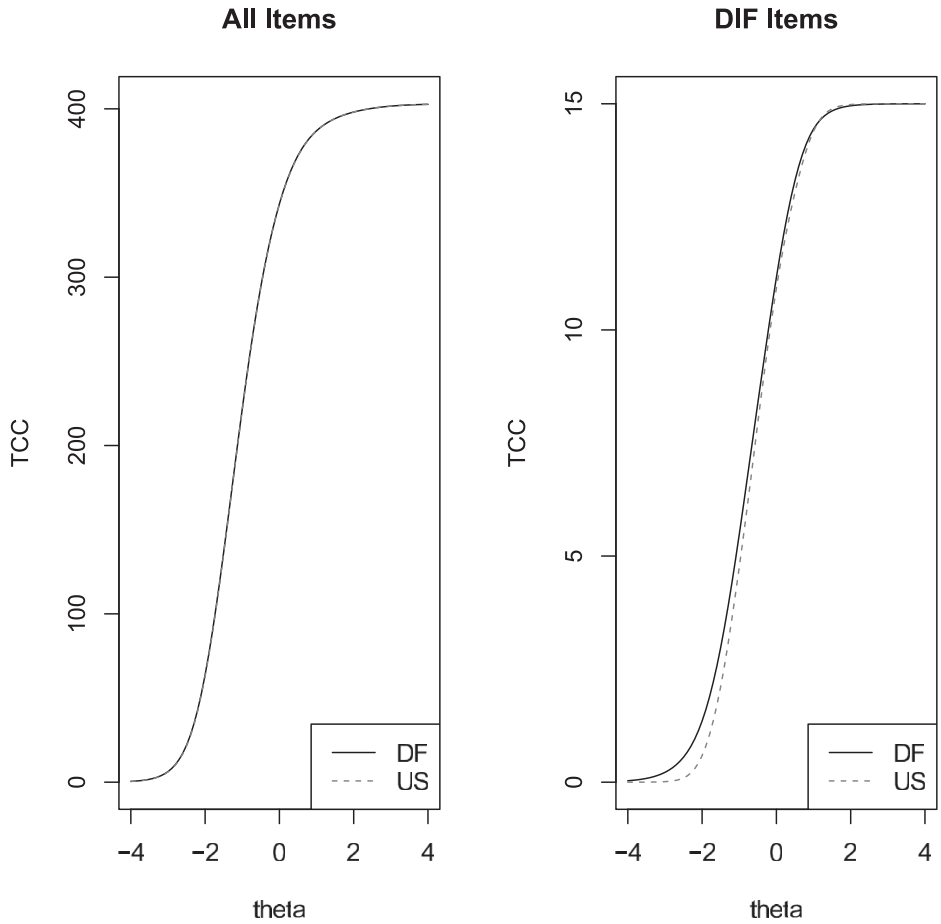

Fig. 1. The overall impact of DIF for language on the test characteristic curves (TCCs). The TCC shows the relation between the total item scores (y-axis) and theta (x-axis). Left graph shows the TCC for all 121 Dutch–Flemish (DF) and US PROMIS Physical Function items (ignoring DIF); the right graph shows the TCC for the four Dutch–Flemish and US PROMIS Physical Function items having DIF. DIF, differential item functioning; PROMIS, Patient-Reported Outcomes Measurement Information System.

PROMIS Physical Function AMS-PAIN sample was 35.7 (SD = 7.4), with a range from 10.2 to 73.5, indicating a lower level of physical function in Dutch chronic pain patients compared with persons from the US general population.

Fig. 2 shows plots of the standard errors across the range of the Dutch–Flemish PROMIS Physical Function T-scores, for the total Dutch-Flemish PROMIS Physical Function item bank (121 items), the 4-, 6-, 8-, 10-, and 20-item Short Forms, and CATs respectively. The results indicate good reliability of the Dutch–Flemish PROMIS Physical Function item bank. The reliability of the total item bank was greater than 0.90 for the full range of the scale presented in Fig. 2. Both the 10- and 20-item Short Forms as well as all CATs show a reliability of 0.80 or greater between a T-score of 25 and 48, where 90% of the Dutch AMS-PAIN sample is located. The 10- and 20-item CATs show even greater reliability than 0.90 across this T-score range. The plots demonstrate that CATs show greater reliability than the Short Forms. For example, where a 20-item short form is needed for persons with T-scores between 30 and 50 to achieve a reliability of greater than 0.90, a CAT can achieve the same level of reliability with only 10 items.

3.6. Construct validity

The Dutch–Flemish PROMIS Physical Function item bank correlated strongly (>-0.70) with the Dutch–Flemish PROMIS Pain Interference item bank, as expected ($r = -0.73$). In addition, the Dutch–Flemish PROMIS Physical Function item bank correlated strongly as expected ($r < -0.50$) with the disability instruments (NDI, −0.70; DASH, −0.86; RMDQ, −0.70; and FIQ, −0.62). The correlation with the Dutch-Flemish PROMIS Global Health Pain intensity item was higher than expected: −0.62 instead of between −0.30 and −0.50.

4. Discussion

The results of this study add to the evidence on the psychometric properties of the Dutch-Flemish PROMIS Physical Function item bank by showing unidimensionality, good model fit, and breadth of coverage across the range of physical function. Furthermore, the analyses showed no evidence for DIF due to gender and administration mode, negligible DIF by age, and good cross-cultural validity, reliability, and construct validity.

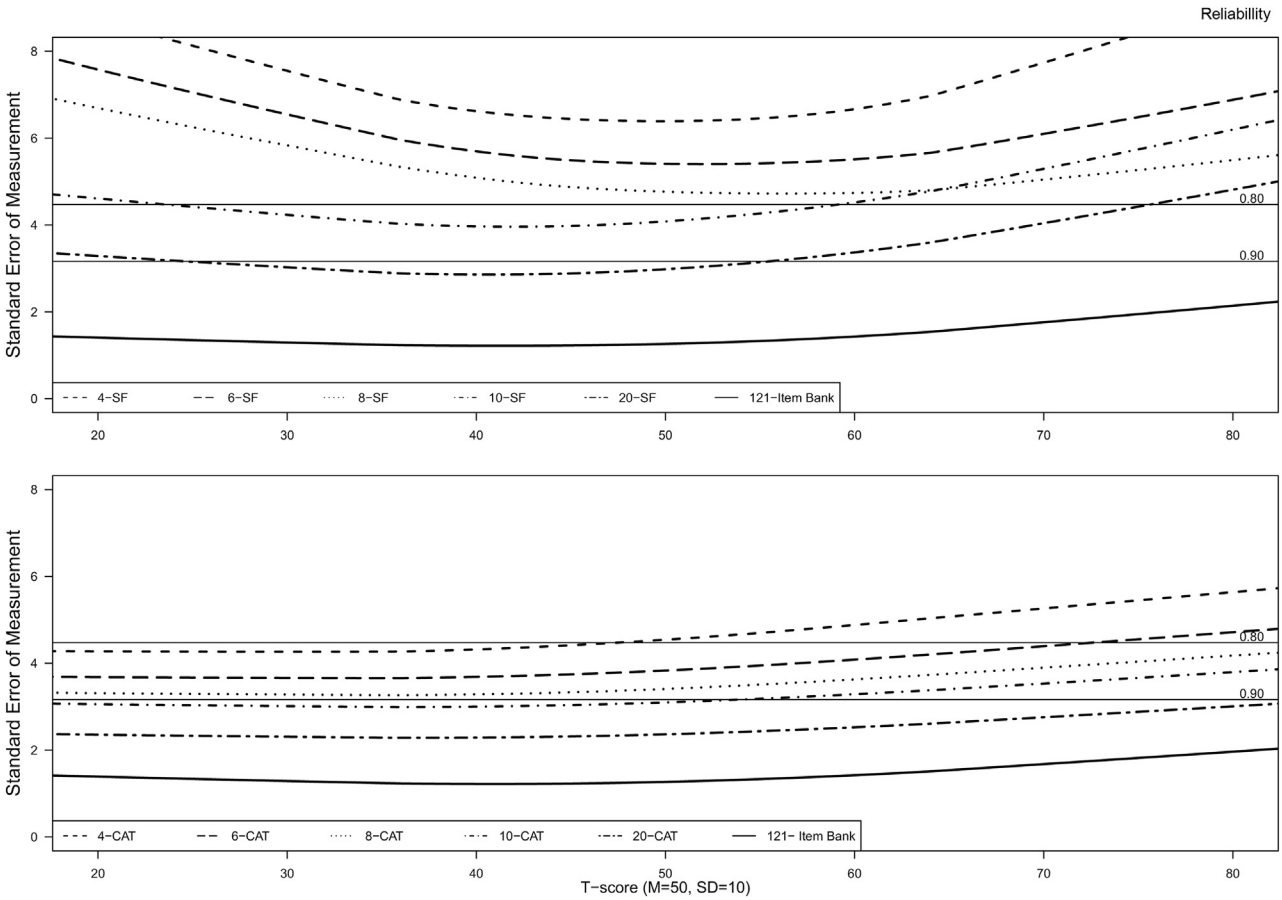

Fig. 2. The plots show the standard errors of the total Dutch-Flemish PROMIS Physical Function item bank (121-item bank) and the 4-, 6-, 8-, 10-, and 20-item Short Forms (SFs) and CATs, respectively. The horizontal axis represents the different physical function abilities with T = 50 representing the mean of the US general population with a standard deviation of 10. The vertical axis represents the standard error (reliability), with reference reliabilities of 0.80 and 0.90. The lower the curve, the greater the reliability. CAT, computerized adaptive testing; M, mean; PROMIS, Patient-Reported Outcomes Measurement Information System; SD, standard deviation.

The Dutch–Flemish PROMIS Group aims to improve the measurement of patient-reported outcomes in the Netherlands and Flanders by providing and supporting the implementation of IRT-based, efficient, highly reliable, and valid PROMIS items banks and CATs [23]. PROMIS item banks are based on well-developed conceptual models with clearly defined unidimensional constructs and have been developed by extensive qualitative research with patients [20]. PROMIS item banks show good measurement properties and are more applicable for use in daily clinical practice than traditional PROMs [18–20]. Through the use of IRT-based methods, PROMIS T-scores can be estimated even when people do not respond to the same items, for instance when using CAT. CATs are tailored to the persons' ability and therefore more efficient and precise than other PROMs [21,22]. The PROMIS scores are expressed on a standardized T-score metric (T-score 50 represents the average score of the general US population, with a standard deviation of 10) that facilitates interpretability of scores by providing a consistent reference point [21,22]. Recently, Schalet et al. published cross-walk tables that can be used to convert scores of the traditional physical function PROMs, such as the SF-36 PF and the HAQ-DI, into PROMIS T-scores [69–72]. With use of the cross-walk tables, historical data can be mapped onto the PROMIS T-score metric, making it possible to switch from using traditional physical function PROMs to PROMIS measures [69].

Although the response rate in this study was only 46%, the large sample size of 1,247 patients is reassuring. The low response rate might be due to the high number of items that were administered to the patients.

No items were flagged for DIF with respect to gender or administration mode, and there was negligible DIF by age, which means that the items and scores can be used across patients that differ in gender and age, and differ in the way of completing the item bank (digital or paper). This is reassuring compared with previous research with the Dutch–Flemish PROMIS Physical Function item bank, in which seven items were flagged for DIF with respect to gender and five with respect to age in Dutch rheumatoid arthritis patients [25].

Evaluation of cross-cultural validity (Dutch–Flemish vs. US) flagged only four of the 121 items with DIF for

language. Caution is warranted in interpreting this DIF for language because of other differences between the Dutch and US populations. The Dutch AMS-PAIN sample differs from the US wave 1 sample in age, gender, pain intensity, and educational level. These differences are likely explained by the fact that the Dutch sample comprises chronic pain patients (with large proportion female and higher pain intensities as common characteristics), compared to persons from the general population in the US sample. However, because of the minimal impact of DIF on the item scores in present study, we conclude that the cross-cultural item differences were negligible and that all items can be retained in the item bank for the time being. Three items with DIF for language had also been found to have DIF in a previous study of the Dutch−Flemish PROMIS Physical Function item bank, in which 25 items were flagged for DIF for language in RA patients [25]. The DIF item PFB1 is included in the PROMIS Physical Function short forms. This could possibly affect the comparability of the US and Dutch T-scores resulting from short forms. For this DIF item, we think translational improvement may be possible. Changing the wording of this item may decrease the DIF for language. We recommend testing new (possibly better) translation of this item in a future data collection.

Because of similar calibration properties and the negligible DIF between the Dutch−Flemish and the US PROMIS Physical function item bank found in present study and also in the study of Oude Voshaar et al. and because of the value of using common calibrations for purposes of international comparability, we recommend that US PROMIS Physical Function item parameters be used in the Dutch−Flemish Physical Function CAT. In future studies, it would be interesting to evaluate the impact of DIF on the Dutch−Flemish PROMIS Physical Function scores obtained by CAT, by comparing CAT T-scores applying the Dutch−Flemish item parameters with CAT T-scores applying the original US item parameters. The impact of DIF may be greater when using CAT compared with using the total item bank because it would be theoretically possible for the CAT to select only DIF items [50]. However, this is unlikely, and can its probability could be lessened by ensuring that no DIF-flagged item was used as the starting item.

The present study partly supports the construct validity of the Dutch−Flemish PROMIS Physical Function item bank. The Dutch−Flemish PROMIS Physical Function item bank scores had a high negative correlation with scores on the Dutch−Flemish PROMIS Pain Interference item bank (−0.73). In addition, the correlations between the Dutch−Flemish PROMIS Physical Function item bank and the disability instruments were high, as hypothesized. However, we expect slightly lower correlations with the disability instruments than with the Pain Interference item bank because the disability instruments measure the concept of disability more broadly, but this was not confirmed. Furthermore, the correlation between the Physical Function item bank and the PROMIS Pain Intensity item was higher than expected (−0.62 instead of between −0.30 and −0.50), which does not support the discriminant validity. However, a high correlation between pain and physical function instruments has often been found in the literature [73].

For future research, we recommend investigating whether the Physical Function and Pain Interference item banks are measuring one single construct in patients with chronic pain. Furthermore, we recommend evaluating the test−retest reliability of the Dutch−Flemish PROMIS Physical Function item bank.

In conclusion, this study found that Dutch−Flemish PROMIS Physical Function item bank responses had good fit to the GRM, substantial cross-cultural validity, and good reliability, across the continuum of physical function, and construct validity. The item bank has the potential to improve the measurement of physical function across a wide range of populations. The Dutch−Flemish PROMIS Physical Function item bank and short forms are now available for clinical application in Dutch-speaking persons and a Dutch−Flemish PROMIS Physical Function CAT can now be developed, for the time being with US PROMIS Physical Function item parameters. The Dutch−Flemish PROMIS Physical Function item bank and short forms can be obtained through the web site www.dutchflemishpromis.nl.

Acknowledgments

The Dutch−Flemish PROMIS group is an initiative that aims to translate and implement PROMIS item banks and CATs in the Netherlands and Flanders (www.dutchflemishpromis.nl). The authors would like to thank Kiki Dirix, Jacqueline Bruinsma, and all employees of the movement laboratory and logistics department of Reade (Centre for Rehabilitation and Rheumatology in the Netherlands) for all their administrative support.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2017.03.011.

References

[1] PROMIS. NIH PROMIS website 2015. Available at http://www.nihpromis.org. Accessed July 20, 2015.
[2] Harstall C, Ospina M. How prevalent is chronic pain. Pain Clin Update 2003;11:1−4.
[3] Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:3−19.
[4] Paterson DH, Warburton DE. Physical activity and functional limitations in older adults: a systematic review related to Canada's Physical Activity Guidelines. Int J Behav Nutr Phys Act 2010;7:38.
[5] Oude Voshaar MA, ten Klooster PM, Taal E, van de Laar MA. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health Qual Life Outcomes 2011;9:99.

[6] Kocks JWH, Asijee GM, Tsiligianni IG, Kerstjens HA, van der Molen T. Functional status measurement in COPD: a review of available methods and their feasibility in primary care. Prim Care Respir J 2011;20:269−75.
[7] Jordhoy MS, Ringdal GI, Helbostad JL, Oldervoll L, Loge JH, Kaasa S. Assessing physical functioning: a systematic review of quality of life measures developed for use in palliative care. Palliat Med 2007;21:673−82.
[8] Grotle M, Brox JI, Vøllestad NK. Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. Spine (Phila Pa 1976) 2005;30: 130−40.
[9] Barten JA, Pisters MF, Huisman PA, Takken T, Veenhof C. Measurement properties of patient-specific instruments measuring physical function. J Clin Epidemiol 2012;65:590−601.
[10] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45:S3−11.
[11] Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks. Med Care 2007;45:22−31.
[12] Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63:1179−94.
[13] Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 2014;67:516−26.
[14] Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2008;61:17−33.
[15] Gershon R. Computer adaptive testing. J Appl Meas 2005;6: 109−27.
[16] Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
[17] Cella D, Gershon R, Lai J, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007;16:133−41.
[18] Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol 2011;38:1759−64.
[19] Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther 2011;13:R147.
[20] Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, Hays RD. The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). Arthritis Care Res (Hoboken) 2011; 63(Suppl 1):S486−90.
[21] Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061−6.
[22] Lai JS, Cella D, Choi S, Junghaenel DU, Christodoulou C, Gershon R. How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS fatigue item bank example. Arch Phys Med Rehabil 2011;92:20−7.
[23] Terwee CB, Roorda LD, de Vet HCW, Dekker J, Westhovens R, van Leeuwen J, et al. Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). Qual Life Res 2014;23:1733−41.
[24] Haverman L, Grootenhuis MA, Raat H, van Rossum MA, van Dulmen-den Broeder E, Hoppenbrouwers K, et al. Dutch−Flemish translation of nine pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®. Qual Life Res 2016;25:761−5.
[25] Oude Voshaar MAH, ten Klooster PM, Glas CAW, Vonkeman HE, Taal E, Krishnan E, et al. Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. PLoS One 2014;9:e92367.
[26] Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Eval Health Prof 2005;28:212−32.
[27] Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11:R191.
[28] Amtmann D, Cook KF, Jensen MP, Chen W-H, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain 2010;150:173−82.
[29] Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration and validation of the Dutch-Flemish PROMIS pain interference item bank in patients with chronic pain. PLoS One 2015;10:e0134094.
[30] Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009;18:873−80.
[31] Sendlbeck M, Araujo E, Schett G, Englbrecht M. Psychometric properties of three single-item pain scales in patients with rheumatoid arthritis seen during routine clinical care: a comparative perspective on construct validity, reproducibility and internal responsiveness. RMD Open 2015;1:e000140.
[32] Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409−15.
[33] Vos CJ, Verhagen AP, Koes BW. Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. Eur Spine J 2006;15:1729−36.
[34] Ailliet L, Rubinstein SM, de Vet HCW, van Tulder MW, Terwee CB. Reliability, responsiveness and interpretability of the neck disability index-Dutch version in primary care. Eur Spine J 2015;24:88−93.
[35] Jorritsma W, de Vries GE, Dijkstra PU, Geertzen JHB, Reneman MF. Neck Pain and Disability Scale and Neck Disability Index: validity of Dutch language versions. Eur Spine J 2012;21: 93−100.
[36] Köke A, Heuts P, Vlaeyen J. Neck Disability Index (NDI). In: Centrum PK, editor. Meetinstrumenten chronische pijn. Maastricht: Pijn Kennis Centrum, Academisch Ziekenhuis Maastricht; 1996: 52−4.
[37] Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med 1996;29:602−8.
[38] Palmen C, Van der Meijden E, Nelissen Y, Köke A. De betrouwbaarheid en validiteit van de Nederlandse vertaling van de Disability of the Arm, Shoulder, and Hand questionnaire (DASH). Ned Tijdschr Voor Fysiother 2004;114:30−5.
[39] Bot SDM, Terwee CB, van der Windt DAWM, Bouter LM, Dekker J, de Vet HCW. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis 2004;63:335−41.
[40] Huisstede BMA, Feleus A, Bierma-Zeinstra SM, Verhaar JA, Koes BW. Is the disability of arm, shoulder, and hand questionnaire (DASH) also valid and responsive in patients with neck complaints. Spine (Phila Pa 1976) 2009;34:E130−8.
[41] Veehof MM, Sleegers EJA, van Veldhoven NHMJ, Schuurman AH, van Meeteren NLU. Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). J Hand Ther 2002;15: 347−54.

[42] Gommans IHB, Koes BW, Van Tulder MW. Validiteit en responsiviteit Nederlandstalige Roland Disability Questionnaire. Vragenlijst naar functionele status bij patienten met lage rugpijn. Ned Tijdschr Voor Fysiother 1997;107:28−33.
[43] Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976) 1983;8:141−4.
[44] Beurskens AJ, de Vet HC, Köke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71−6.
[45] Brouwer S, Kuijer W, Dijkstra PU, Göeken LNH, Groothoff JW, Geertzen JHB. Reliability and stability of the Roland Morris Disability Questionnaire: intra class correlation and limits of agreement. Disabil Rehabil 2004;26:162−5.
[46] Köke A, Heuts P, Vlaeyen J. Roland Disability Questionnaire. In: Centrum PK, editor. Meetinstrumenten chronische pijn. Maastricht: Pijn Kennis Centrum, Academisch Ziekenhuis Maastricht; 1996:68−70.
[47] Burckhardt CS, Clark SR, Bennett RM. The fibromyalgia impact questionnaire: development and validation. J Rheumatol 1991;18:728−33.
[48] Zijlstra TR, Taal E, van de Laar MAFJ, Rasker JJ. Validation of a Dutch translation of the fibromyalgia impact questionnaire. Rheumatology (Oxford) 2007;46:131−4.
[49] Köke A, Heuts P, Vlaeyen J. Fibromyalgia Impact Questionnaire. In: Centrum PK, editor. Meetinstrumenten chronische pijn. Maastricht: Pijn Kennis Centrum, Academisch Ziekenhuis Maastricht; 1996:36−8.
[50] Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45:S22−31.
[51] Crins MHP, Roorda LD, Smits N, de Vet HCW, Westhovens R, Cella D, et al. Calibration of the Dutch-Flemish PROMIS pain behavior item bank in patients with chronic pain. Eur J Pain 2016;20:284−96.
[52] Rosseel Y. Lavaan: an R package for structural equation modeling. J Stat Softw 2012;48:1−36.
[53] R-Software. Available at. www.r-project.org. Accessed June 3, 2014.
[54] Reckase M. Unifactor latent trait models applied to multifactor tests: results and implications. J Educ Stat 1979;4:207−30.
[55] Mokken RJ. A theory and procedure of scale analysis: with applications in political research. The Hague: Mouton; 1971.
[56] Sijtsma K, Molenaar I. Introduction to nonparametric item response theory. Thousand Oaks: Sage; 2002.
[57] Van der Ark L. Mokken scale analysis in R. J Stat Softw 2007;20:1−19.
[58] Sijtsma K, Emons WHM, Bouwmeester S, Nyklícek I, Roorda LD. Nonparametric IRT analysis of quality-of-life scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). Qual Life Res 2008;17:275−90.
[59] Chalmers RP. The mirt package: multidimensional item response theory. Library of the R package; 2016. Available at www.cran.r-project.org/web/packages/mirt/mirt.pdf.
[60] Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Softw 2012;48:1−29.
[61] McKinley R, Mills C. A comparison of several goodness-of-fit statistics. Appl Psychol Meas 1985;9:49−57.
[62] Holland P, Wainer H. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993.
[63] Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Med Care 2006;44:S115−23.
[64] Choi SW, Gibbons LE, Crane PK. Logistic ordinal regression differential Item Functioning using IRT, version 0.3-3 n.d. Available at https://cran.r-project.org/web/packages/lordif/index.html.
[65] Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw 2011;39:1−30.
[66] Revicki DA, Chen W-H, Harnam N, Cook KF, Amtmann D, Callahan LF, et al. Development and psychometric analysis of the PROMIS pain behavior item bank. Pain 2009;146:158−69.
[67] Magis D, Raiche G. Random generation of response patterns under computerized adaptive testing with the R package catR. J Stat Softw 2012;48:1−31.
[68] Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. Educ Psychol Meas 2013;73:5−26.
[69] Schalet BD, Revicki DA, Cook KF, Krishnan E, Fries JF, Cella D. Establishing a common metric for physical function: linking the HAQ-DI and SF-36 PF subscale to PROMIS(®) physical function. J Gen Intern Med 2015;30:1517−23.
[70] ProsettaStone. Available at. www.prosettastone.org. Accessed July 30, 2015.
[71] Choi SW, Podrabsky T, McKinney N, Schalet BD, Cook KF, Cella DIn: PROsetta Stone® Methodology A Rosetta Stone for Patient Reported Outcomes, 1 2015. Available at http://www.prosettastone.org/Methodology/Documents/PROSetta%20Methodology%20Report.pdf.
[72] Choi SW, Podrabsky T, McKinney N, Schalet BD, Cook KF, Cella DIn: PROsetta Stone® Analysis Report A Rosetta Stone for Patient Reported Outcomes, 1 2015. Available at http://www.prosettastone.org/AnalysisReport/Documents/PROsetta%20Stone%20Analysis%20Report_Vol%202_09_15-2016.pdf.
[73] Collins N, Prinsen C, Christensen R, Bartels E, Terwee C, Roos E. Knee Injury and Osteoarthritis Outcome Score (KOOS): systematic review and meta-analysis of measurement properties. Osteoarthr Cartil 2016;24:1317−29.