Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank in patients with chronic pain

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

DOI
10.1002/ejp.727

Publication date
2016

Document Version
Final published version

Published in
European Journal of Pain

EJP
European Journal of Pain

ORIGINAL ARTICLE

Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank in patients with chronic pain

M.H.P. Crins[1], L.D. Roorda[1], N. Smits[2], H.C.W. de Vet[3], R. Westhovens[4,5], D. Cella[6], K.F. Cook[6], D. Revicki[7], J. van Leeuwen[8], M. Boers[3,9], J. Dekker[10], C.B. Terwee[3]

1 Amsterdam Rehabilitation Research Center, Reade, Amsterdam, The Netherlands
2 Department of Clinical Psychology and Department of Methodology, The EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands
3 Department of Epidemiology and Biostatistics, The EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands
4 Department of Development and Regeneration, Skeletal Biology and Engineering Research Center, KU Leuven, Louvain, Belgium
5 Rheumatology, University Hospitals, KU Leuven, Louvain, Belgium
6 Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, USA
7 Outcomes Research, Evidera, Bethesda, USA
8 Leones Group BV, Amsterdam, The Netherlands
9 Department of Rheumatology, VU University Medical Center, Amsterdam, The Netherlands
10 Department of Rehabilitation Medicine and Department of Psychiatry, VU University Medical Center, Amsterdam, The Netherlands

Correspondence
M.H.P. Crins
E-mail: m.crins@reade.nl

Funding sources
The Dutch-Flemish translation of the PROMIS item banks was supported by a grant from the Dutch Arthritis Association. Funding for the current calibration study was provided by a grant from the Dutch Scientific College of Physical Therapy.

Conflicts of interest
None declared.

Accepted for publication
12 April 2015

doi:10.1002/ejp.727

Abstract

Background: The aims of the current study were to calibrate the item parameters of the Dutch-Flemish PROMIS Pain Behavior item bank using a sample of Dutch patients with chronic pain and to evaluate cross-cultural validity between the Dutch-Flemish and the US PROMIS Pain Behavior item banks. Furthermore, reliability and construct validity of the Dutch-Flemish PROMIS Pain Behavior item bank were evaluated.
Methods: The 39 items in the bank were completed by 1042 Dutch patients with chronic pain. To evaluate unidimensionality, a one-factor confirmatory factor analysis (CFA) was performed. A graded response model (GRM) was used to calibrate the items. To evaluate cross-cultural validity, Differential item functioning (DIF) for language (Dutch vs. English) was evaluated. Reliability of the item bank was also examined and construct validity was studied using several legacy instruments, e.g. the Roland Morris Disability Questionnaire.
Results: CFA supported the unidimensionality of the Dutch-Flemish PROMIS Pain Behavior item bank (CFI = 0.960, TLI = 0.958), the data also fit the GRM, and demonstrated good coverage across the pain behavior construct (threshold parameters range: −3.42 to 3.54). Analysis showed good cross-cultural validity (only six DIF items), reliability (Cronbach's α = 0.95) and construct validity (all correlations ≥0.53).
Conclusions: The Dutch-Flemish PROMIS Pain Behavior item bank was found to have good cross-cultural validity, reliability and construct validity. The development of the Dutch-Flemish PROMIS Pain Behavior item bank will serve as the basis for Dutch-Flemish PROMIS short forms and computer adaptive testing (CAT).

What's already known about this topic
- PROMIS instruments are becoming a gold standard in clinical care and scientific research for measuring patient-reported health.

What does this study add
- Calibrating the Dutch-Flemish translation of the PROMIS Pain Behavior item bank.
- This study will serve as the basis for Dutch-Flemish PROMIS short forms and computer adaptive testing, which contribute to tailor-made measurement of patient-reported health in, e.g. chronic pain patients.

1. Introduction

Chronic pain is defined as pain that persists beyond the normal tissue healing time (Harstall and Ospina, 2003). The prevalence of chronic pain in western populations is high (10.1–55.2%) (Picavet and Schouten, 2003; Cimmino et al., 2011; Reid et al., 2011). The most prevalent pain is musculoskeletal pain, with prevalence varying from 30 to 40% for low back pain, 15–20% for shoulder and neck pain, 10–15% for chronic widespread pain and 2% for fibromyalgia (Harstall and Ospina, 2003; Picavet and Schouten, 2003). Chronic pain often leads to substantial limitations in daily activities (Harstall and Ospina, 2003).

The term, pain behaviors, refers to behaviors of people with pain that communicate their pain experiences to others (Fordyce, 1976; Keefe et al., 2001). These behaviors may include facial expressions, verbal complaints, gestures, and postural adjustments, taking medication or resting. Besides expressions of pain, it has been suggested that pain intensity and pain-related impairments decreases when pain behaviors decrease (Fordyce et al., 1986). Consequently, pain behaviors are useful treatment targets and important to measure in people with chronic pain (Hadjistavropoulos et al., 2007; Turk et al., 2008). One way of gaining insight into pain behavior is by use of Patient-Reported Outcome Measures (PROMs). However, the availability of PROMs for pain behavior is scant (Revicki et al., 2009).

The National Institutes of Health (NIH) Patient-Reported Outcomes Measurement Information System (PROMIS®) initiative has developed a new assessment system for measuring patient-reported health, including an item bank for measuring pain behavior (Cella et al., 2007, 2010; Reeve et al., 2007a). An item bank is a set of questions (items) measuring the same construct, e.g. pain behavior. Items from existing PROMs were collected, combined and transformed into the PROMIS Pain Behavior item bank (Revicki et al., 2009). Furthermore, new items were developed and added to the item bank. The PROMIS Pain Behavior item bank was calibrated using item response theory (IRT) methods and can be used as a tailor-made measure of patient-reported health, for instance via a computerized adaptive test (CAT) (Reeve et al., 2007b). A CAT is a computer-administered test in which the successive items are chosen based on answers given to previous questions. Because the administration of items is tailored, individuals respond to a minimal number of relevant questions. PROMIS item banks and CATs have the potential to be implemented worldwide; they have been shown to have strong content validity, responsiveness and other desirable psychometric properties (Fries et al., 2011a,b; Khanna et al., 2011; Magasi et al., 2012). Furthermore, PROMIS scores are expressed on a standardized *T*-score metric and therefore easier to interpret than traditional PROMs scores. The Dutch-Flemish PROMIS Group translated 17 adult PROMIS item banks and 9 paediatric PROMIS item banks into Dutch-Flemish (the Dutch-speaking part of Belgium), including the PROMIS Pain Behavior item bank (Terwee et al., 2014).

The first aim of the current study was to calibrate the Dutch-Flemish PROMIS Pain Behavior item bank in Dutch chronic pain patients. The second aim was to evaluate cross-cultural validity of the Dutch-Flemish PROMIS Pain Behavior item bank compared to the United States (US) PROMIS Pain Behavior item bank. The third aim was to evaluate the reliability and construct validity of the Dutch-Flemish PROMIS Pain Behavior item bank.

2. Methods

2.1 Study participants

For this study, 2808 patients from the Amsterdam Pain (AMS-PAIN) cohort were invited to participate. The AMS-PAIN cohort consists of chronic pain patients who have been registered since September 2010 in Reade; an outpatient secondary care centre for rheumatology and rehabilitation in the Netherlands. To be eligible, patients had to have at least one chronic pain condition for at least 3 months prior to 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 versus the US PROMIS Pain Behavior item bank, data from the US PROMIS Pain Behavior American Chronic Pain Association (ACPA) sample were used. The ACPA sample consists of 967 patients with chronic pain who responded to 31 of the 39 items from the PROMIS Pain Behavior item bank (Revicki et al., 2009). Because of potential response burden issues, eight items from the PROMIS Pain Behavior item bank were not administered to the ACPA sample. All ACPA chronic pain patients met the study eligibility criteria of being 21 years or older and having at least one chronic pain condition for at least 3 months prior to participating in the US PROMIS Wave 1 study (Revicki et al., 2009).

2.2 Procedures

Patients from the AMS-PAIN cohort were invited by e-mail or letter, to fill in a web-based (digital) or paper-and-pencil (paper) questionnaire that included, among other measures, the full Dutch-Flemish PROMIS Pain Behavior item bank. For the digital questionnaire, patients received personal login codes. Patients who were unable to complete the digital questionnaire were asked to complete the paper version. The study was approved by the local institutional review board (of Slotervaart and Reade).

2.3 Measures

The questionnaire included the full Dutch-Flemish PROMIS Pain Behavior item bank. The translation of the US PROMIS Pain Behavior item bank into Dutch-Flemish was performed by Functional Assessment of Chronic Illness Therapy multilingual translation (FACITtrans) using standardized methodology and approved by the PROMIS Statistical Center (Eremenco et al., 2005; Terwee et al., 2014). This translation included multiple forward and back translations, independent reviews and pilot testing with cognitive debriefing among 70 Dutch and Flemish adults (Terwee et al., 2014). The Dutch-Flemish PROMIS Pain Behavior item bank contains 39 items covering a wide range of pain behaviors (Revicki et al., 2009). Patients have to rate how frequently they engage in the pain behaviors in the past 7 days, using a 6-point Likert response scale (1 = Had no pain, 2 = Never, 3 = Rarely, 4 = Sometimes, 5 = Often, 6 = Always). Demographic information also was collected (i.e. age, gender, country of birth, educational level).

In addition, the questionnaire contained five legacy instruments including the pain intensity item (Global07) from the Dutch-Flemish PROMIS Global Health item bank (an 11-point numeric rating scale (NRS) with 0 = 'no pain' and 10 = 'worst pain imaginable') (Hays et al., 2009). Four condition-specific instruments were included, which are widely accepted and have demonstrated reliability and validity. The Neck Disability Index (NDI) was used for patients with chronic neck pain. 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 in which higher scores indicate more disability (Vernon and Mior, 1991; Vos et al., 2006). Evidence has accumulated for the reliability and validity of the NDI within Dutch patients with chronic neck pain (Köke et al., 1996a; Vos et al., 2006; Jorritsma et al., 2012; Ailliet et al., 2014). The Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire was used for patients with chronic shoulder pain. The DASH consists of 30 items measuring disabilities of the upper extremities, with a total score ranging from 0 to 100 in which higher scores indicate more disability (Hudak et al., 1996; Palmen et al., 2004). DASH scores have demonstrated good reliability and validity in Dutch patients with a variety of unilateral disorders of the upper limb (Veehof et al., 2002; Bot et al., 2004; Palmen et al., 2004; Huisstede et al., 2009). The Roland Morris Disability Questionnaire (RMDQ) was used for patients with chronic back pain. The RMDQ consists of 24 items measuring disabilities as a result of chronic back pain, with a total score ranging from 0 to 24. Higher scores indicate more disability (Roland and Morris, 1983; Gommans et al., 1997). RMDQ scores have demonstrated good reliability and validity within Dutch patients with chronic low back pain (Beurskens et al., 1996; Köke et al., 1996b; Gommans et al., 1997; Brouwer et al., 2004). The fourth condition-specific legacy instrument was the Fibromyalgia Impact Questionnaire (FIQ), used for patients with fibromyalgia. The FIQ consists of 20 items measuring physical disabilities as a result of fibromyalgia, with a total score ranging from 0 to 100 (Burckhardt et al., 1991; Zijlstra et al., 2007). Higher scores indicate higher impact of fibromyalgia on physical ability (Burckhardt et al., 1991; Zijlstra et al., 2007). FIQ scores have demonstrated moderate to good reliability and validity among Dutch patients with fibromyalgia (Köke et al., 1996c; Zijlstra et al., 2007).

2.4 Statistical analysis

2.4.1 Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank

The psychometric analyses were conducted using the PROMIS analysis plan (Reeve et al., 2007b). To evaluate the measurement properties of the pain behavior items, IRT-based analyses were used. IRT models estimate the relationship between an item response category and the level of the measured construct, in this study the level of pain behavior. Before calibrating the item parameters of the Dutch-Flemish PROMIS Pain Behavior item bank, the three IRT assumptions, unidimensionality, local independence and monotonicity, were evaluated (Reeve et al., 2007b).

Unidimensionality was examined using confirmatory factor analyses (CFA) in which all items were hypothesized to load on a single factor. The analysis was performed using 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) (Rosseel, 2012; R-Software, 2014). The criteria for unidimensionality include CFI >0.95, TLI >0.95 and RMSEA <0.06 (Reeve et al., 2007b). Furthermore, unidimensionality was considered sufficient when 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 (Reckase, 1979; Reeve et al., 2007b). This was examined with exploratory factor analysis (EFA).

Another IRT assumption is local independence, which means that after controlling for the dominant factor, there should be no significant covariance among item responses. Local dependency was evaluated by examining the residual correlation matrix resulting from the single factor CFA. Residual correlations greater than 0.2 were considered as indicators of possible local dependence (Reeve et al., 2007b). In addition, local independence was studied using Yen's Q3 statistic (Yen, 1993). This statistic calculates the residual item scores under the graded response model (GRM) and correlates these among items. Cohen's rules of thumb were used for correlation effect sizes (Cohen, 1988). In this, Q3 values between 0.24 and 0.36 are moderate deviations, and values of 0.37 and greater represent large deviations. Items that show local dependence were examined and their impact on IRT parameters estimates was evaluated, by removing the locally dependent items one by one and examining changes in the IRT parameters of the remaining items (Reeve et al., 2007b).

A third IRT assumption is monotonicity, in which 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 Pain Behavior items was evaluated by fitting a non-parametric IRT model, using Mokken scaling in the R package Mokken (Mokken, 1971; Sijtsma and Molenaar, 2002; Van der Ark, 2007; Sijtsma et al., 2008). This model yields non-parametric IRT response curve estimates, shows the probabilities of endorsing response categories and can be visually inspected to evaluate monotonicity.

After evaluation of the IRT assumptions, a GRM was fit to the data to calibrate the item parameters using Multilog version 7 (Thissen et al., 2003). The GRM models two item parameters, the item threshold and the item slope (Reeve et al., 2007b). Item threshold parameters indicate the item difficulty, and locate the items along the measured trait. The item slope parameter represents the discriminative ability of the items, with higher slope values indicating better ability to discriminate between adjoining values on the construct.

To assess the fit of the GRM and the degree in which possible misfit affects the IRT model, S-X^2 statistic was used (McKinley and Mills, 1985). 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. Items that show a S-X^2 p-value of less than 0.001, demonstrate poor fit (McKinley and Mills, 1985; Reeve et al., 2007b).

2.4.2 Differential item functioning

Differential item functioning (DIF) analyses are used to examine if people from different groups (e.g. age or gender) with the same level of trait (in this study the same level of pain behavior) have different probabilities of giving a certain response to an item (Holland and Wainer, 1993; Embretson and Reise, 2000; Reeve et al., 2007b). There are two kinds of DIF: uniform and non-uniform (Holland and Wainer, 1993; Embretson and Reise, 2000; Reeve et al., 2007b). Uniform DIF exists when the DIF is consistent, with the same magnitude of DIF across the entire range of the trait. Non-uniform DIF exists when the magnitude or direction of DIF differs across the trait. DIF was evaluated with use of the R package Lordif (version 0.2-2), using ordinal logistic

regression models with a McFadden's pseudo R^2 change of 2% as critical value (Crane et al., 2006; Reeve et al., 2007b; Choi et al., 2011). 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).

2.4.3 Cross-cultural validity

Differences in descriptive characteristics between the Dutch AMS-PAIN patients and the US ACPA patients were evaluated with use of independent sample *t*-tests and χ^2 tests, for continuous and categorical variables, respectively.

Cross-cultural validity of the Dutch-Flemish PROMIS Pain Behavior item bank versus the US PROMIS Pain Behavior item bank was assessed to examine if Dutch and US patients with the same level of trait have different probabilities of giving a certain response to the translated Dutch-Flemish item (Dutch patients) or the original English item (US patients), respectively. This addresses the comparability of the Dutch-Flemish and US PROMIS Pain Behavior scores. For the evaluation of cross-cultural validity, DIF for language (Dutch vs. English) was analysed with use of the R package Lordif (version 0.2-2), using ordinal logistic regression models with a McFadden's pseudo R^2 change of 2% as critical value (Crane et al., 2006; Reeve et al., 2007b; Choi et al., 2011). In this, DIF was evaluated for 31 out of the 39 items, because the items PAINBE32, -34, -38, -40, -41, -46, -47 and -48 were missing in the US ACPA sample (Revicki et al., 2009). When items were flagged as potential DIF items, the impact of DIF was examined by plotting item characteristic curves (ICC) (data not shown) and test characteristic curves (TCC). The TCC plots showed the scores for all 31 Pain Behavior items (ignoring DIF), and the scores for only the items having DIF (Choi et al., 2011).

2.4.4 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 no or mild pain behavior to high levels of pain behavior) (Revicki et al., 2009; Amtmann et al., 2010). The formula indicates that increased scale information is related to smaller SEs and, therefore, greater measurement precision. Using the calculated SEs, plots were overlaid showing standard errors (as a parameter of reliability) across the score range of the total Dutch-Flemish PROMIS Pain Behavior item bank, the 7-item short form (v1.0.7a, similar to the US short form) and a 7-item simulated CAT. For this, the IRT theta scores of the Dutch AMS-PAIN sample were transformed into *T*-scores anchored on the US item parameters of the PROMIS Pain Behavior item bank which can be found in the US PROMIS Pain Behavior cue sheet; the spreadsheet listing the items, response options and item parameters of the US PROMIS Pain Behavior item bank (Revicki et al., 2009). This cue sheet can be found on the US PROMIS website www.nihpromis.org. The *T*-score 50 represents the average score of the general US population, with a standard deviation of 10.

2.4.5 Construct validity

Construct validity of the Dutch-Flemish PROMIS Pain Behavior item bank was evaluated by correlating the *T*-scores of the Dutch-Flemish PROMIS Pain Behavior item bank to the scores on the legacy instruments, including the Dutch-Flemish PROMIS Global Health pain intensity item score and the total scores of the NDI, DASH, RMDQ and FIQ. Construct validity was evaluated using Pearson's correlations. Because the NDI, DASH, RMDQ and FIQ are disability measures rather than pain impact measures, we hypothesized that Dutch-Flemish PROMIS Pain Behavior item bank scores would have the highest correlation ($r > 0.50$) with the Dutch-Flemish PROMIS Global Health pain intensity item and lower correlations ($r = 0.30$–0.50) with other legacy instruments.

3. Results

3.1 Demographic characteristics

Of the 2808 invited patients of the Dutch AMS-PAIN cohort, 1140 responded to the questionnaire (response rate 40.6%). No differences were found between responders and non-responders on age, gender, country of birth or education level. Among the 1140 respondents, 29 patients were excluded because they did not give informed consent. IRT analyses showed that the first response category 'had no pain' negatively affected the IRT parameter estimations. It was decided to delete the response

category 'had no pain', resulting in IRT analysis with five response options (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always). Furthermore, patients who reported having no pain were excluded (n = 69). Of the net sample of 1042 patients, 929 had complete data. The statistical analyses were performed with these 929. However, because the GRM analyses can accommodate incomplete data, all 1042 were used for the IRT calibration.

The demographic characteristics of the Dutch AMS-PAIN sample and the US ACPA chronic pain sample are summarized in Table 1. Of the AMS-PAIN patients, 79% (n = 822) were women and the average age (SD) was 48 years (12) with a range from 21 to 85. Eighty-five percent (n = 608) of these were born in the Netherlands, and 82% had at least a high school degree. Of the AMS-PAIN patients, 82% (n = 855) indicated that the duration of their pain was more than 2 years, and the average pain intensity on a NRS(SD) was 6.7 (1.9). Of all patients, 71% reported that they had chronic low back pain, 70% had chronic neck or shoulder pain, 36% reported fibromyalgia, 47% chronic widespread pain, 35% reported migraine or other chronic headache and 35% reported osteoarthritis. Twelve percent of the chronic pain patients reported that they had rheumatoid arthritis and less than 2% reported cancer. No differences were found in age, gender and pain intensity between the Dutch AMS-PAIN sample and the US ACPA sample. However, there were some differences in educational levels, pain duration and type of chronic pain condition. The Dutch AMS-PAIN patients reported slightly more often pain duration of 1–2 years. The Dutch AMS-PAIN sample was less educated with 32% reporting high school education or less, while the US ACPA sample reported 19% high school education level or less.

3.2 Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank

The CFA results indicated good fit to a unidimensional model. The CFI was 0.960 and the TLI was 0.958, which are above the criterion of >0.95 (Reeve et al., 2007b). However, the RMSEA was 0.099, which is somewhat larger than the criterion of <0.06. The first factor in EFA accounted for 42% of the variance, and the second factor accounted for 6% of the variance; hence, the ratio of the variance explained by the first to the second factor is 7, which is higher than the required minimum of 4 (Reeve et al., 2007b). Based on these results, it was concluded that the Dutch-Flemish PROMIS Pain Behavior items share a single common factor, and are sufficiently unidimensional.

Examining the residual correlation matrix showed a small number of deviances from local dependence among the items. Fourteen out of the 741 item pairs (1.9%) were marked as possibly locally dependent. The results of the GRM using Yen's Q3 statistic show that 22 of the 741 item pairs (3%) had Q3 values that showed at least a moderate deviation of model fit. The item pairs with the greatest dependency were PAINBE18 *("When I was in pain I asked for help doing things that needed to be done")* – PAINBE25 *("When I was in pain I called out for someone to help me")* with residual correlation of 0.56, and PAINBE42 *("When I was in pain my upper body would tense up")* – PAINBE47 *("When I was in pain my body became stiff")* with residual correlation of 0.53. These items were removed one by one and then the impact on item parameters of the remaining items was evaluated. Mean differences in item thresholds after item removal ranged from 0.004 to 0.03, and the maximum absolute differences ranged from 0.01 to 0.10. The mean differences in the slope parameter of the remaining items ranged from 0.006 to 0.01, and the maximum absolute differences in the slope parameter ranged from 0.03 to 0.05. These results suggest that the impact of local dependence was minimal. Therefore, no items were removed from the item bank because of local dependency.

Evaluation of monotonicity showed that the Pain Behavior item bank complied with monotonicity to a great extent; only a few mild and non-significant violations were found. The Mokken scalability coefficient of the whole pain behavior scale was 0.38, in concordance with a scale of weak quality according to rules of thumb (Mokken, 1971; Van der Ark, 2007). One out of the 39 items (PAINBE41: *"When I was in pain I screamed"*) had a scalability coefficient of 0.29, which was slightly lower than the lower bound of 0.30. Based on these results, it was concluded that the Dutch-Flemish PROMIS Pain Behavior items sufficiently met the assumption of monotonicity.

Table 2 summarizes the IRT item parameters of the Dutch-Flemish PROMIS Pain Behavior items. The item slope parameters ranged from 1.00 to 2.35. The item with lowest discrimination parameter was PAINBE29 *("When I was in pain I used a cane or something else for support")*, and the item with the highest discrimination parameter was PAINBE28 *("When I was in pain I squirmed")*. The item threshold parameters ranged from −3.42 to 3.54.

Table 1 Demographic characteristics of the Dutch AMS-PAIN sample (n = 1042) and the US ACPA sample (n = 967).

	Dutch chronic pain sample	US chronic pain sample
Age mean (SD) range	48 (12) 21–85	48 (11) 21–86
Gender *n* (%)		
Male	220 (21)	182 (19)
Female	822 (79)	780 (81)
Country of birth *n* (%)		
Netherlands	608 (58)	–
Other	434 (42)	–
Social status[a] *n* (%)		
Single	361 (35)	–
Married or living together	550 (53)	–
Living apart together	55 (5)	–
Living with parents	22 (2)	–
Other	61 (6)	–
Educational level *n* (%)		
Less than High School degree	177 (18)	23 (3)***
High School degree	141 (14)	157 (16)
Some college	393 (41)	452 (47)
College degree	44 (5)	214 (22)***
Advanced degree	211 (22)	114 (12)***
Employment status[a] *n* (%)		
Full-time	163 (16)	–
Part-time	265 (25)	–
Student	42 (4)	–
Unpaid, volunteer, household	161 (16)	–
Retired	82 (8)	–
Unemployed	182 (18)	–
Other	213 (20)	–
Social benefits[a] *n* (%)		
Sick listed	236 (23)	–
Disability benefit	240 (23)	–
Unemployment benefit	80 (8)	–
Other	132 (13)	–
No social benefit	361 (35)	–
Duration of pain *n* (%)		
3–6 months	14 (1)	15 (2)
6–12 months	34 (3)	38 (4)
1–2 years	132 (13)	65 (7)***
2–5 years	300 (29)	234 (24)
>5 years	555 (53)	577 (60)
Type of chronic pain condition[a] *n* (%)		
Migraine and/or other 'daily' headache	369 (35)	209 (22)***
Rheumatoid arthritis	127 (12)	59 (6)***
Osteoarthritis	367 (35)	195 (20)***
Pain related to cancer	19 (2)	8 (.8)*
Lower back pain	744 (71)	533 (55)***
Neck or shoulder pain	730 (70)	447 (46)***
Fibromyalgia	370 (36)	338 (35)
Chronic widespread pain	489 (47)	–
Other neuropathic pain (nerve damage)	208 (20)	370 (38)***
Other	475 (46)	298 (31)***
No chronic pain condition	3 (.3)	1 (.1)***
T-score mean (SD)	60.7 (4.1)	63.7 (3.5)
range	50.8–90.0	54.0–78.6***
Legacy instruments mean (SD)		
PROMIS Global Health Pain intensity	6.7 (1.9)	6.6 (1.6)
NDI	25.0 (9)	–
DASH	46.0 (20)	–
RMDQ	13.0 (6)	–
FIQ	60.0 (17)	–

PROMIS Global Health Pain Intensity (0–10); NDI, Neck Disability Index (0–50); DASH, Disabilities of the Arm, Shoulder and Hand (0–100); RMDQ, Roland Morris Disability Questionnaire (0–24); FIQ, Fibromyalgia Impact Questionnaire (0–100).

[a]Multiple answers were allowed.

*$p < 0.05$.

***$p < 0.001$.

The probability values for the S-X^2 statistics ranged from 0.0000 to 0.8176. Based on the S-X^2 p-value of less than 0.001, only 3 of 39 items were found to misfit the GRM. From all calibration results, it was concluded not to remove items from the item bank.

3.3 Differential item functioning

None of the Dutch-Flemish PROMIS Pain Behavior items were flagged for DIF for age, gender or administration mode.

3.4 Cross-cultural validity

The analysis of DIF for language, flagged six items with some level of DIF (see Table 2). Five were flagged for uniform DIF. For three out of these five items (PAINBE22 (R^2 = 0.039): *"Pain caused me to bend over while walking"*, PAINBE25 (R^2 = 0.060): *"When I was in pain I called out for someone to help me"* and PAINBE26 (R^2 = 0.48): *"Pain caused me to curl up in a ball"*), the Dutch patients were more likely to endorse higher response categories compared to the US patients who were at the same level of the trait. For two out of these five items (PAINBE42 (R^2 = 0.023): *"When I was in pain my upper body would tense up"* and PAINBE50 (R^2 = 0.039): *"When I was in pain I moved my limbs protectively"*), the Dutch patients were more likely to endorse lower response

Table 2 IRT item characteristics for the Dutch-Flemish PROMIS Pain Behavior item bank.

Item code	Item	Slope	Category threshold				Mokken's	Item fit statistics	
		a	B1	B2	B3	B4	H	S-X^2	Prob X^2
PAINBE2	When I was in pain I became irritable	1.18	−3.42	−2.12	−0.12	1.82	0.336	67.44	0.0043
PAINBE3	When I was in pain I grimaced[b]	1.39	−2.37	−1.48	0.16	2.04	0.349	43.79	0.3137
PAINBE6	When I was in pain I would lie down	1.16	−2.33	−1.09	0.43	2.26	0.310	73.84	0.0009***
PAINBE8	When I was in pain I moved extremely slowly	1.52	−2.35	−1.31	−0.13	1.31	0.387	34.74	0.7055
PAINBE9	When I was in pain I became angry	1.56	−1.33	−0.41	0.93	2.14	0.382	40.74	0.4374
PAINBE11	When I was in pain I clenched my teeth	1.49	−1.17	−0.33	0.80	2.22	0.366	46.04	0.2364
PAINBE13	When I was in pain I tried to stay very still	1.60	−1.58	−0.53	0.66	2.05	0.394	46.32	0.2277
PAINBE16	When I was in pain I appeared upset or sad	2.04	−1.28	−0.43	0.64	1.69	0.434	46.81	0.2130
PAINBE17	When I was in pain I gasped	1.88	−0.59	0.15	1.25	2.53	0.425	44.18	0.2993
PAINBE18	When I was in pain I asked for help doing things that needed to be done	1.11	−2.14	−0.84	0.66	2.38	0.306	88.75	0.0000***
PAINBE21	When I was in pain it showed on my face (squinching eyes, opening eyes wide, frowning)	1.84	−1.52	−0.77	0.32	1.52	0.413	31.84	0.8176
PAINBE22	Pain caused me to bend over while walking[a] α	1.65	−1.06	−0.49	0.60	1.78	0.402	39.84	0.1990
PAINBE23	When I was in pain I asked one or more people to leave me alone	1.55	−1.20	−0.44	0.88	2.11	0.379	47.30	0.1990
PAINBE24	When I was in pain I moved stiffly	1.11	−3.02	−1.81	−0.18	2.19	0.318	43.14	0.3386
PAINBE25	When I was in pain I called out for someone to help me[a] α	1.24	−1.27	−0.12	1.49	2.85	0.329	92.77	0.0000***
PAINBE26	Pain caused me to curl up in a ball[a] α	2.25	−0.99	−0.32	0.70	1.85	0.456	64.07	0.0092
PAINBE27	I had pain so bad it made me cry	1.84	−0.94	−0.36	0.91	2.09	0.408	50.24	0.1287
PAINBE28	When I was in pain I squirmed	2.35	−0.79	−0.13	0.85	1.96	0.459	51.72	0.1014
PAINBE29	When I was in pain I used a cane or something else for support	1.00	0.17	0.85	1.81	2.99	0.286	38.79	0.5247
PAINBE31	I limped because of pain	1.24	−1.67	−0.94	0.27	1.72	0.334	34.04	0.7347
PAINBE32	When I was in pain I became quiet and withdrawn	1.56	−1.99	−1.22	0.12	1.78	0.395	45.30	0.2606
PAINBE33	When I was in pain I frowned	1.64	−1.15	−0.38	0.80	2.14	0.395	52.00	0.0968
PAINBE34	When I was in pain I asked for help when walking or changing positions	1.46	−0.20	0.46	1.52	2.71	0.396	45.46	0.2552
PAINBE35	When I was in pain I groaned	1.78	−1.07	−0.28	0.94	2.16	0.399	53.65	0.0731
PAINBE37	When I was in pain I isolated myself from others	1.46	−1.47	−0.68	0.67	2.21	0.374	47.92	0.1824
PAINBE38	When I was in pain I drew my knees up	1.26	−0.48	0.24	1.52	3.03	0.333	66.50	0.0053
PAINBE39	When I was in pain I moaned, whined or whimpered	1.82	−0.66	0.20	1.19	2.38	0.407	38.23	0.5504
PAINBE40	When I was in pain I flung my arms or limbs around	1.10	0.45	1.25	2.50	3.54	0.312	42.85	0.3498
PAINBE41	When I was in pain I screamed	1.95	0.16	0.81	1.75	2.61	0.448	51.39	0.1070
PAINBE42	When I was in pain my upper body would tense up[a] β	1.34	−1.09	−0.42	0.68	2.21	0.356	44.22	0.2979
PAINBE43	When I was in pain I walked carefully	1.40	−2.21	−1.37	−0.21	1.15	0.371	41.98	0.3853
PAINBE44	When I was in pain I bit or pursed my lips	1.60	−0.46	0.16	1.25	2.65	0.394	50.28	0.1279
PAINBE45	When I was in pain I thrashed	1.61	0.37	1.12	2.00	3.03	0.410	56.15	0.0465
PAINBE46	When I was in pain I protected the part of my body that hurt	1.54	−1.25	−0.57	0.44	1.55	0.372	41.11	0.4219
PAINBE47	When I was in pain my body became stiff	1.31	−1.70	−0.87	0.41	1.94	0.349	33.98	0.7370
PAINBE48	When I was in pain I clenched my jaw or gritted my teeth	1.51	−0.65	0.06	1.04	2.17	0.371	44.01	0.3055
PAINBE49	When I was in pain I winced	1.64	−1.46	−0.67	0.54	2.08	0.389	60.69	0.0190
PAINBE50	When I was in pain I moved my limbs protectively[a] β	1.64	−0.62	0.12	1.11	2.39	0.392	52.63	0.0870
PAINBE51	When I was in pain I avoided physical contact with others	1.46	−1.41	−0.66	0.31	1.66	0.366	57.41	0.0366

[a]uniform DIF; α activity is relatively faster endorsed in the Dutch chronic pain patients; β activity is relatively faster endorsed in US chronic pain patients.

[b]non-uniform DIF.

***$p < 0.001$.

categories compared to the US patients. One item showed non-uniform DIF (PAINBE3 ($R^2 = 0.025$): *"When I was in pain I grimaced"*).

The overall impact of DIF for language on the TCC is shown in Fig. 1. The left graph shows the TCC for all 31 Pain Behavior items (ignoring DIF), and the right graph shows the TCC for just the six items having DIF. These curves show that the Pain Behavior total score is only slightly higher for Dutch patients than for US patients, indicating minimal impact of DIF by language.

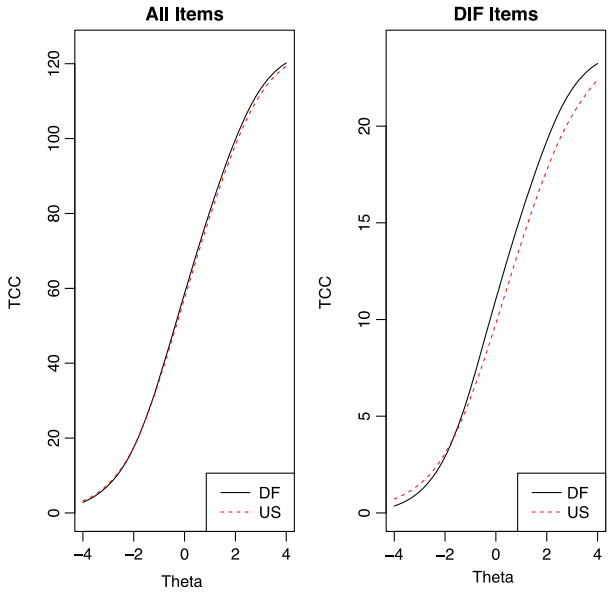

Figure 1 The overall impact of DIF for language on the test characteristic curves (TCC). The TCC shows the relation between the total item scores (y-axis) and theta (x-axis). Left graph shows the TCC for all 31 Dutch-Flemish (DF) and United States (US) PROMIS Pain Behavior items (ignoring DIF); the right graph shows the TCC for just the 6 items having DIF.

3.5 Reliability

As shown in Table 1, the mean *T*-score for the overall Dutch-Flemish PROMIS Pain Behavior AMS-PAIN sample was 60.7 (SD = 4.1), with a range from 50.8 to 90.0. 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 cue sheet of the US PROMIS Pain Behavior item bank. Fig. 2 shows the distribution of the Dutch AMS-PAIN (Dutch clinical) sample and the US Wave1 (US general population) and ACPA (US clinical) sample along the *T*-score scale. Fig. 2 shows plots of the standard errors across the range of the Dutch-Flemish PROMIS Pain Behavior *T*-scores, for the total item bank, the 7-item PROMIS short form (v1.0.7a) and a 7-item simulated CAT. The reliability of the total item bank was 0.90 or greater across most of the *T*-score distribution. The 7-item short form and the 7-item simulated CAT have reliabilities above 0.80 across the *T*-score distribution. The Cronbach's alpha estimate for the total item bank was 0.95. These results indicate good reliability of the Dutch-Flemish PROMIS Pain Behavior item bank.

3.6 Construct validity

Pearson correlation coefficients, shown in Table 3, indicate the relations between *T*-scores of the Dutch-Flemish PROMIS Pain Behavior item bank and the total scores on the legacy instruments. The correlation between scores on the Dutch-Flemish PROMIS Pain Behavior item bank and on the PROMIS Global Health Pain intensity item was 0.56

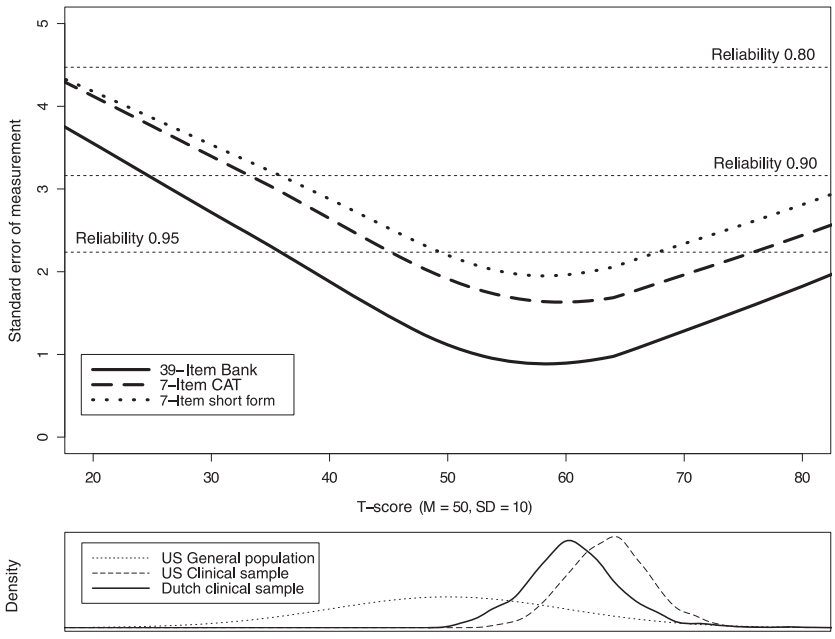

Figure 2 Standard error of measurement of the Dutch-Flemish PROMIS Pain Behavior item bank.

Table 3 Correlations between the Dutch-Flemish PROMIS Pain Behavior *T*-scores and the legacy instruments.

Instrument	*N*	Expected *R*	Observed *R*
PROMIS Global Health Pain intensity	872	>0.50	0.56
NDI	379	0.30–0.50	0.60
DASH	375	0.30–0.50	0.63
RMDQ	626	0.30–0.50	0.61
FIQ	283	0.30–0.50	0.53

NDI, Neck Disability Index; DASH, Disabilities of the Arm, Shoulder and Hand; RMDQ, Roland Morris Disability Questionnaire; FIQ, Fibromyalgia Impact Questionnaire.

($p < 0.001$), which is above 0.50 as expected. The Dutch-Flemish PROMIS Pain Behavior item bank correlated higher than expected (0.30–0.50) with all legacy instruments, with correlations between 0.53 and 0.63.

4. Discussion

The aim of current study was to calibrate the Dutch-Flemish PROMIS Pain Behavior item bank in Dutch patients with chronic pain and to evaluate cross-cultural validity of the Dutch-Flemish versus the US PROMIS Pain Behavior item bank. The construct validity of the Dutch-Flemish PROMIS Pain Behavior item bank also was evaluated. The results of this study support the Dutch-Flemish PROMIS Pain Behavior item bank's unidimensionality, model fit and breadth of coverage across the range of pain behavior. Furthermore, the analyses of the Dutch-Flemish PROMIS Pain Behavior item bank showed no evidence for DIF due to age, gender, and administration mode, good cross-cultural validity, reliability and construct validity. This study is the first calibration study of the Dutch-Flemish PROMIS Pain Behavior item bank.

The Dutch-Flemish PROMIS Group aims to improve the measurement of patient-reported outcomes in the Netherlands and Flanders (the Dutch-speaking part of Belgium) by providing and supporting the implementation of efficient, IRT-based, highly reliable and valid PROMIS items banks and CATs (Terwee et al., 2014). PROMIS item banks and CATs have better content validity compared to existing traditional PROMs (Magasi et al., 2012). PROMIS item banks are based on a well-developed conceptual model with clearly defined unidimensional constructs, and have been developed using extensive qualitative research with patients (Khanna et al., 2011). PROMIS item banks show good measurement properties; they have small measurement errors and show better responsiveness compared to more traditional PROMs (Fries et al., 2011a,b; Khanna et al., 2011). This makes the use of PROMIS item banks in daily clinical practice more suitable than traditional PROMs. Increased responsiveness results in reduction in sample sizes needed in clinical studies (Fries et al., 2011a,b). Through the use of IRT-based methods, PROMIS item bank and CAT scores approximate an interval scale instead of an ordinal scale, and therefore are easier to interpret than scores of more traditional PROMs (Fries et al., 2009; Lai et al., 2011). The PROMIS scores are expressed on a common standardized *T*-score metric, and because they are calibrated using an IRT model, the *T*-scores can be estimated even if people do not respond to the same items, for instance when using CAT. The use of CAT has great advantages compared to more traditional paper questionnaires; CATs are more efficient and have better precision (Fries et al., 2009; Lai et al., 2011).

The analyses of the IRT assumptions show that the required assumptions of unidimensionality, local independence and monotonicity are met. The CFA results, CFI as well as the TLI supported unidimensionality. The RMSEA was beyond the criterion of <0.06, but RMSEA values tend to be elevated when the number of items is large (Cook et al., 2009).

The calibration analyses of the Dutch-Flemish PROMIS Pain Behavior items show that the range of the item threshold parameters indicates good coverage across the range of the pain behavior construct. Furthermore, the item threshold parameters show which items are most useful for measuring different levels of Pain Behavior, which is required for the selection process of items in a CAT.

No items were flagged for DIF with respect to gender, age and administration mode. Therefore, the Dutch-Flemish PROMIS Pain Behavior items and scores can be used across patients who differ in gender and age, and differ in the way of completing the item bank (digital or paper).

Although the response rate in this study was only 40.2%, the large sample size of 1042 patients is reassuring. When comparing the Dutch AMS-PAIN sample with the US ACPA sample, no differences were found in age, gender and pain intensity. However, the differences in educational level are noteworthy, where the Dutch AMS-PAIN patients were less educated than the US ACPA patients (32% vs. 19% reporting high school education or less).

The evaluation of cross-cultural validity of the Dutch-Flemish PROMIS Pain Behavior item bank versus the US PROMIS Pain Behavior item bank

identified evidence of DIF for language across 6 of 31 items used for these DIF analyses. However, DIF had a minimal impact on the item scores. Therefore, we conclude that the cross-cultural item differences were negligible and that all items can be retained in the item bank. Furthermore, we recommend that US PROMIS Pain Behavior item parameters can be used for the time being for the development of the Dutch-Flemish Pain Behavior CAT. To implement a CAT in clinical practice and scientific research, it is necessary to obtain (e.g. from US PROMIS) or develop CAT software and to translate or develop a CAT user interface. Furthermore, there is likely a need for a country-specific server and website, where the Dutch-Flemish CATs can be administered and where the data can be stored.

This study supports the construct validity of the Dutch-Flemish PROMIS Pain Behavior item bank, in which the correlations between the Dutch-Flemish PROMIS Pain Behavior item bank and the legacy instruments were higher than hypothesized. As expected, the Dutch-Flemish PROMIS Pain Behavior item bank correlated highly with the Dutch-Flemish PROMIS Global Health pain intensity item (0.56). However, the correlations with the other legacy instruments were higher than expected (0.53–0.61). Although these legacy instruments are not actually measuring pain but measuring level of disability, other research has similarly reported moderately high correlations between instruments measuring pain and those measuring disability (Gommans et al., 1997; Beaton et al., 2001; Jorritsma et al., 2012). Further, self-reported disability as a result of pain is similar conceptually to self-reported pain behavior, possibly contributing to the high correlations between pain behavior and level of disability.

The Dutch-Flemish PROMIS Pain Behavior item bank is ready to be used as an item bank or short form within Dutch persons with chronic pain. A standard 7-item short form was developed by PROMIS (v1.0.7a), including items with the highest information value. When selecting Dutch-Flemish PROMIS Pain Behavior items for short forms, it would be preferable to select items without DIF. However, two items (PAINBE3 and PAINBE25) that show DIF for language are included in the Pain Behavior 7-item short form. This could possibly affect the comparability of the US and Dutch scores resulting from the 7-item short form. For both items there are some potential translational improvements. Therefore, we recommend testing new (possibly better) translations of the items that exhibited DIF in a future data collection. Nevertheless, the 7-item Dutch-Flemish PROMIS Pain Behavior short form, which is equal to the US PROMIS Pain Behavior short form (v1.0.7a), can be used because the impact of DIF is likely to be small.

The Dutch-Flemish PROMIS Pain Behavior item bank is now calibrated and ready for use by Dutch persons with chronic pain. For the time being, we recommend to use US PROMIS Pain Behavior item parameters and the US *T*-score metric, with $T = 50$ as mean *T*-score of the general US population as reference point and on which the Dutch chronic pain sample is anchored with a mean *T*-score of 60.7. We recommend future analyses on data collected with the Dutch-Flemish PROMIS Pain Behavior item bank in the general Dutch and Flemish population, Flemish patients and in patient groups with other health problems resulting in (chronic) pain. After data collection in the general Dutch and Flemish population, the item bank needs to be recalibrated, and then a Dutch-Flemish *T*-score metric can be developed with a $T = 50$ as mean *T*-score of the general Dutch-Flemish population as reference point. Also, it should then be decided if Dutch-Flemish specific item parameters are needed or whether the US item parameters can also be used in Dutch-Flemish patients. Furthermore, for future research it would be interesting to study DIF for other factors than age, gender, administration mode and language. Besides, it would be interesting to evaluate the impact of DIF on the Dutch-Flemish PROMIS Pain Behavior scores obtained by CAT, by comparing a CAT applying the Dutch-Flemish item parameters with a CAT applying the original US item parameters. The impact of DIF may be greater when using CAT as compared to using the total item bank because a CAT uses only a small item set (Reeve et al., 2007b). Another important step for future research and also for implementing the Dutch-Flemish PROMIS Pain Behavior item bank, short forms and CAT, is to further improve the interpretability of the PROMIS metric. For example, by estimating PROMIS Pain Behavior score differences that represent clinical severity thresholds and clinically meaningful category intervals, also called minimal important changes. Recently, Cook et al.(2013) developed the Pain Behaviors Self Report (PaB-SR) based on the PROMIS Pain Behavior item bank (Cook et al., 2013). Although the PaB-SR is not a PROMIS instrument, future research should consider if the adjustments in the PaB-SR are also relevant for the (Dutch-Flemish) PROMIS instrument.

In conclusion, this item calibration study found good cross-cultural and construct validity of the Dutch-Flemish PROMIS Pain Behavior item bank. The item bank has the potential to improve the measurement of pain behavior. The Dutch-Flemish PROMIS Pain Behavior item bank and short form are now available for clinical application in Dutch-speaking persons with chronic pain and a Dutch-Flemish PROMIS Pain Behavior CAT can now be developed, for the time being using US PROMIS Pain Behavior item parameters. The full Dutch-Flemish PROMIS Pain Behavior item bank and short form (including instructions, item context and response options) are available at the website www.dutchflemishpromis.nl.

Author contributions

All authors were responsible for study concept and design, critical revision of manuscript for important intellectual content and final approval of the version of the manuscript submitted. M.C. was responsible for acquisition of data. M.C, L.R., N.S., C.T. were responsible for statistical analysis and interpretation of data and drafting of manuscript.

Acknowledgements

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). We thank Kiki Dirix 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.

References

Ailliet, L., Rubinstein, S.M., de Vet, H.C.W., van Tulder, M.W., Terwee, C.B. (2015). Reliability, responsiveness and interpretability of the neck disability index-Dutch version in primary care. *Eur Spine J* 24, 88–93.

Amtmann, D., Cook, K.F., Jensen, M.P., Chen, W.-H., Choi, S., Revicki, D., Cella, D., Rothrock, N., Keefe, F., Callahan, L., Lai, J.-S. (2010). Development of a PROMIS item bank to measure pain interference. *Pain* 150, 173–182.

Van der Ark, L. (2007). Mokken Scale Analysis in R. *J Stat Softw* 20, 1–19.

Beaton, D.E., Katz, J.N., Fossel, A.H., Wright, J.G., Tarasuk, V., Bombardier, C. (2001). Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. *J Hand Ther* 14, 128–146.

Beurskens, A.J., de Vet, H.C., Köke, A.J. (1996). Responsiveness of functional status in low back pain: A comparison of different instruments. *Pain* 65, 71–76.

Bot, S.D.M., Terwee, C.B., van der Windt, D.A.W.M., Bouter, L.M., Dekker, J., de Vet, H.C.W. (2004). Clinimetric evaluation of shoulder disability questionnaires: A systematic review of the literature. *Ann Rheum Dis* 63, 335–341.

Brouwer, S., Kuijer, W., Dijkstra, P.U., Göeken, L.N.H., Groothoff, J.W., Geertzen, J.H.B. (2004). Reliability and stability of the Roland Morris Disability Questionnaire: Intra class correlation and limits of agreement. *Disabil Rehabil* 26, 162–165.

Burckhardt, C.S., Clark, S.R., Bennett, R.M. (1991). The fibromyalgia impact questionnaire: Development and validation. *J Rheumatol* 18, 728–733.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J.F., Gershon, R., Hahn, E.A., Lai, J.-S., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol* 63, 1179–1194.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J.F., Bruce, B., Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Med Care* 45, S3–S11.

Choi, S.W., Gibbons, L.E., Crane, P.K. (2011). 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* 39, 1–30.

Cimmino, M.A., Ferrone, C., Cutolo, M. (2011). Epidemiology of chronic musculoskeletal pain. *Best Pract Res Clin Rheumatol* 25, 173–183.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (New Jersey: Lawrence Erlbaum Associates).

Cook, K.F., Kallen, M.A., Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Qual Life Res* 18, 447–460.

Cook, K.F., Keefe, F., Jensen, M.P., Roddey, T.S., Callahan, L.F., Revicki, D., Bamer, A.M., Kim, J., Chung, H., Salem, R., Amtmann, D. (2013). Development and validation of a new self-report measure of pain behaviors. *Pain* 154, 2867–2876.

Crane, P.K., Gibbons, L.E., Jolley, L., van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Med Care* 44, S115–S123.

Embretson, S.E., Reise, S.P. (2000). *Item Response Theory for Psychologists* (Mahwah, NJ: Lawrence Erlbaum).

Eremenco, S.L., Cella, D., Arnold, B.J. (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Eval Health Prof* 28, 212–232.

Fordyce, W.E. (1976). *Behavioral Methods for Chronic Pain and Illness* (St. Louis, MO: C.V. Mosby).

Fordyce, W.E., Brockway, J.A., Bergman, J.A., Spengler, D. (1986). Acute back pain: A control-group comparison of behavioral vs traditional management methods. *J Behav Med* 9, 127–140.

Fries, J., Rose, M., Krishnan, E. (2011a). The PROMIS of better outcome assessment: Responsiveness, floor and ceiling effects, and Internet administration. *J Rheumatol* 38, 1759–1764.

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. *J Rheumatol* 36, 2061–2066.

Fries, J.F., Krishnan, E., Rose, M., Lingala, B., Bruce, B. (2011b). Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Res Ther* 13, R147.

Gommans, I.H.B., Koes, B.W., Van Tulder, M.W. (1997). Validiteit en responsiviteit Nederlandstalige Roland Disability Questionnaire. Vragenlijst naar functionele status bij patienten met lage rugpijn. *Ned Tijdschr Voor Fysiother* 107, 28–33.

Hadjistavropoulos, T., Herr, K., Turk, D.C., Fine, P.G., Dworkin, R.H., Helme, R., Jackson, K., Parmelee, P.A., Rudy, T.E., Lynn Beattie, B., Chibnall, J.T., Craig, K.D., Ferrell, B., Ferrell, B., Fillingim, R.B., Gagliese, L., Gallagher, R., Gibson, S.J., Harrison, E.L., Katz, B.,

Keefe, F.J., Lieber, S.J., Lussier, D., Schmader, K.E., Tait, R.C., Weiner, D.K., Williams, J. (2007). An interdisciplinary expert consensus statement on assessment of pain in older persons. *Clin J Pain* 23, S1–S43.

Harstall, C., Ospina, M. (2003). How prevalent is chronic pain. *Pain Clin Updat* XI, 1–4.

Hays, R.D., Bjorner, J.B., Revicki, D.A., Spritzer, K.L., Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Qual Life Res* 18, 873–880.

Holland, P., Wainer, H. (1993). *Differential Item Functioning* (Hillsdale, NJ: Lawrence Erlbaum Associates).

Hudak, P.L., Amadio, P.C., Bombardier, C. (1996). 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* 29, 602–608.

Huisstede, B.M.A., Feleus, A., Bierma-Zeinstra, S.M., Verhaar, J.A., Koes, B.W. (2009). Is the disability of arm, shoulder, and hand questionnaire (DASH) also valid and responsive in patients with neck complaints. *Spine (Phila Pa 1976)* 34, E130–E138.

Jorritsma, W., de Vries, G.E., Dijkstra, P.U., Geertzen, J.H.B., Reneman, M.F. (2012). Neck Pain and Disability Scale and Neck Disability Index: Validity of Dutch language versions. *Eur Spine J* 21, 93–100.

Keefe, F., Williams, D., Smith, S. (2001). Assessment of pain behaviors. In *Handbook of Pain Assessment*, D. Turk, R. Melzack, eds. (New York, NY: Guilford Press) 170–187.

Khanna, D., Krishnan, E., Dewitt, E.M., Khanna, P.P., Spiegel, B., Hays, R.D. (2011). The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). *Arthritis Care Res (Hoboken)* 63(Suppl 1), S486–S490.

Köke, A., Heuts, P., Vlaeyen, J. (1996a). Neck Disability Index (NDI). In *Meetinstrumenten Chronische Pijn*, Pijn Kennis Centrum, ed. (Maastricht: Pijn Kennis Centrum, Academisch ziekenhuis Maastricht) pp. 52–54.

Köke, A., Heuts, P., Vlaeyen, J. (1996b). Roland Disability Questionnaire. In *Meetinstrumenten Chronische Pijn*, Pijn Kennis Centrum, ed. (Maastricht: Pijn Kennis Centrum, Academisch ziekenhuis Maastricht) pp. 68–70.

Köke, A., Heuts, P., Vlaeyen, J. (1996c). Fibromyalgia Impact Questionnaire. In *Meetinstrumenten Chronische Pijn*, Pijn Kennis Centrum, ed. (Maastricht: Pijn Kennis Centrum, Academisch ziekenhuis Maastricht) pp. 36–38.

Lai, J.S., Cella, D., Choi, S., Junghaenel, D.U., Christodoulou, C., Gershon, R. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Arch Phys Med Rehabil* 92, 20–27.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R.D., Brod, M., Snyder, C., Boers, M., Cella, D. (2012). Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Qual Life Res* 21, 739–746.

McKinley, R., Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Appl Psychol Meas* 9, 49–57.

Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis: With Applications in Political Research* (The Hague: Mouton).

Palmen, C., Van der Meijden, E., Nelissen, Y., Köke, A. (2004). De betrouwbaarheid en validiteit van de Nederlandse vertaling van de Disability of the Arm, Shoulder, and Hand questionnaire (DASH). *Ned Tijdschr Voor Fysiother* 114, 30–35.

Picavet, H.S.J., Schouten, J.S.A.G. (2003). Musculoskeletal pain in the Netherlands: Prevalences, consequences and risk groups, the DMC (3)-study. *Pain* 102, 167–178.

R-Software (2014). http://www.r-project.org/.

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *J Educ Stat* 4, 207–230.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Liu, H., Gershon, R., Reise, S., Lai, J., Cella, D. (2007a). Psychometric evaluation and calibration of health-related quality of life item banks. *Med Care* 45, 22–31.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Liu, H., Gershon, R., Reise, S.P., Lai, J., Cella, D. (2007b). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 45, S22–S31.

Reid, K.J., Harker, J., Bala, M.M., Truyers, C., Kellen, E., Bekkering, G.E., Kleijnen, J. (2011). Epidemiology of chronic non-cancer pain in Europe: Narrative review of prevalence, pain treatments and pain impact. *Curr Med Res Opin* 27, 449–462.

Revicki, D.A., Chen, W.-H., Harnam, N., Cook, K.F., Amtmann, D., Callahan, L.F., Jensen, M.P., Keefe, F.J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain* 146, 158–169.

Roland, M., Morris, R. (1983). 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)* 8, 141–144.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *J Stat Softw* 48, 1–36.

Sijtsma, K., Emons, W.H.M., Bouwmeester, S., Nyklícek, I., Roorda, L.D. (2008). 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* 17, 275–290.

Sijtsma, K., Molenaar, I. (2002). *Introduction to Nonparametric Item Response Theory* (Thousand Oaks: Sage).

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

Thissen, D., Chen, W., Bock, R. (2003). Multilog (version 7). *Sci Softw Int* [computer software].

Turk, D.C., Dworkin, R.H., Revicki, D., Harding, G., Burke, L.B., Cella, D., Cleeland, C.S., Cowan, P., Farrar, J.T., Hertz, S., Max, M.B., Rappaport, B.A. (2008). Identifying important outcome domains for chronic pain clinical trials: An IMMPACT survey of people with pain. *Pain* 137, 276–285.

Veehof, M.M., Sleegers, E.J.A., van Veldhoven, N.H.M.J., Schuurman, A.H., van Meeteren, N.L.U. (2002). Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). *J Hand Ther* 15, 347–354.

Vernon, H., Mior, S. (1991). The Neck Disability Index: A study of reliability and validity. *J Manipulative Physiol Ther* 14, 409–415.

Vos, C.J., Verhagen, A.P., Koes, B.W. (2006). Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. *Eur Spine J* 15, 1729–1736.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *J Educ Meas* 30, 187–213.

Zijlstra, T.R., Taal, E., van de Laar, M.A.F.J., Rasker, J.J. (2007). Validation of a Dutch translation of the fibromyalgia impact questionnaire. *Rheumatology (Oxford)* 46, 131–134.

