Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients With Musculoskeletal Complaints

Schuller, Wouter; Terwee, Caroline B.; Klausch, Thomas; Roorda, Leo D.; Rohrich, Daphne C.; Ostelo, Raymond W.; Terluin, Berend; de Vet, Henrica C.W.

published in
Spine
2019

DOI (link to publisher)
10.1097/BRS.0000000000002847

document version
Publisher's PDF, also known as Version of record

Spine

SPINE Volume 44, Number 6, pp 411–419

EPIDEMIOLOGY

Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients With Musculoskeletal Complaints

Wouter Schuller, MD,[*,†] Caroline B. Terwee, PhD,[*] Thomas Klausch, PhD,[*] Leo D. Roorda, MD, PhD,[‡] Daphne C. Rohrich, MsC,[*] Raymond W. Ostelo, PhD,[*,§] Berend Terluin, MD, PhD,[¶] and Henrica C.W. de Vet, PhD[*]

Study Design. A cross-sectional study.

Objective. The aim of this study was to validate the Dutch-Flemish PROMIS Pain Interference item bank in patients with musculoskeletal complaints.

Summary of Background Data. PROMIS item banks have been developed and validated in the US. They need to be further validated in various patient populations and in different languages.

Methods. One thousand six hundred seventy-seven patients answered the full item bank. A Graded Response Model (GRM) was used to study dimensionality with confirmatory factor analyses and by assessing local independency. Monotonicity was evaluated with Mokken scaling. An Item Response Theory (IRT) model was used to study item fit and to estimate slope and threshold parameters. Differential item functioning (DIF) for language, age, and gender was assessed using ordinal logistic regression analyses. DIF for language was evaluated by comparing our data with a similar US sample. Hypotheses concerning construct validity were tested by correlating item bank-scores with scores on several legacy instruments.

Results. The GRM showed suboptimal evidence of unidimensionality in confirmatory factor analysis [Comparative Fit Index (CFI): 0.903, Tucker-Lewis Index (TLI): 0.897, Root Mean Square Error of Approximation (RMSEA): 0.144], and 99 item pairs with local dependence. A bifactor model showed good fit (CFI: 0.964, TLI: 0.961, RMSEA: 0.089), with a high Omega-H (0.97), a high explained common variance (ECV: 0.81), and no local dependence. Sufficient monotonicity was shown for all items (Mokken $H_{(i)}$: 0.367–0.686). The unidimensional IRT model showed good fit (only two items with S-X^2 < 0.001), with slope parameters ranging from 1.00 to 4.27, and threshold parameters ranging from −1.77 to 3.66. None of the items showed DIF for age or gender. One item showed DIF for language. Correlations with legacy instruments were high (Pearson *R*: 0.53–0.75), supporting construct validity.

Conclusion. The high omega-H and the high ECV indicate that the item bank could be considered essentially unidimensional. The item bank showed good item fit, good coverage of the pain interference trait, and good construct validity.

Key words: headache, item response theory, low back pain, lower extremity pain, musculoskeletal medicine, neck pain, pain interference, PROMIS, upper extremity pain, validity.

Level of Evidence: N/A

Spine 2019;44:411–419

From the [*]VU University Medical Center, Department of Epidemiology & Biostatistics and the Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; [†]Spine Clinic, Zaandam, The Netherlands; [‡]Amsterdam Rehabilitation Research Center, Reade, Amsterdam, The Netherlands; [§]Department of Health Science of the Faculty of Earth and Life Sciences, VU University, Amsterdam, The Netherlands; and [¶]Department of General Practice and Elderly Care Medicine, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands.

Acknowledgment date: May 29, 2018. First revision date: July 26, 2018. Acceptance date: August 3, 2018.

The manuscript submitted does not contain information about medical device(s)/drug(s).

Our study was part of a larger research effort funded by the Dutch Association for Musculoskeletal Medicine.

Relevant financial activities outside the submitted work: board membership, grants.

Address correspondence and reprint requests to Wouter Schuller, MD, VU University Medical Center, Department of Epidemiology & Biostatistics and the Amsterdam Public Health Research Institute, P.O. box 7057, 1007 MB Amsterdam, The Netherlands; E-mail: w.schuller@vumc.nl

DOI: 10.1097/BRS.0000000000002847

In 2004, a US National Institutes of Health (NIH) initiative set out to develop new PROMs for clinical research and health care delivery settings, based upon Item Response Theory (IRT): the Patient Reported Outcome Measurement Information System (PROMIS).[1–3] Under IRT, item banks are constructed consisting of a large collection of questions (*i.e.*, items) covering a wide range of a trait. These item banks are calibrated by modeling the relationship between a person's level of the construct and the likelihood of choosing a response on each item. After calibration, item banks can give comparable scores on a standardized scale, even when subsets of items are used, while retaining reliability.[4–7] Item banks can be used in Computer Adaptive Testing (CAT). In CAT, a computer

algorithm decides on the basis of previous answers which next item would be most informative. The questions are thus tailored to the individual patient and only a small number of questions (on average 5–7) are needed to obtain a reliable score that can be compared with a score obtained from administering all items on the same scale.[3,8–11] IRT outcome measures are expected to play a major role in clinical measurement.[12] The recently suggested research standards from the NIH taskforce for measurement of chronic low back pain, for example, already contain several items from the PROMIS Pain Interference item bank.[13]

The Pain Interference item bank was developed as a unidimensional instrument, measuring the self-reported consequences of pain on relevant aspects of one's life. This includes the extent to which pain hinders engagement with social, cognitive, emotional, physical, and recreational activities. A large number of PROMIS item banks have been translated into Dutch-Flemish by the Dutch-Flemish PROMIS group,[14] among others the v1.1 Pain Interference item bank. A previous study showed good cross-cultural and construct validity, good reliability, and good coverage of the pain interference continuum for the Dutch-Flemish translation of the v1.1 Pain Interference item bank (DF-PROMIS-PI) in a population of rehabilitation patients.[15] For patients with musculoskeletal complaints in the Netherlands, there is a possibility to consult physicians who are trained in musculoskeletal (MSK) medicine.[16] Most MSK practices are primary care facilities primarily focused on patients with MSK pain. Patients generally consult MSK physicians with complaints of low back pain, with or without sciatica, neck pain, headache, and pain in the upper or lower extremities.[16] Before using item banks in patients with various conditions, it is necessary to validate them in different patient populations. For international use, it is necessary to validate item banks in different languages. The aim of our present study was to validate the v1.1 DF-PROMIS-PI item bank in a large sample of patients presented in MSK practice.

MATERIALS AND METHODS

Study Design and Procedure

We conducted a cross-sectional study using an existing web-based registry of patients presenting for the first time in MSK practice. The data documented in this registry were collected by a group of 31 MSK physicians in the Netherlands who agreed to participate in the establishment of the patient registry. At the first visit, the treating physician entered the following patient characteristics *via* computer: age, gender, type and duration of the main complaint, and the existence of concomitant complaints. Complaints were recorded by the treating physician according to the International Classification of Primary Care (ICPC).[17] Treating physicians asked patients whether they were interested in participating in the study. Following an informed consent procedure, the treating physician entered email addresses of the recruited patients in the registry. Thereafter, a specially designed computer program (Readmail) automatically distributed invitations to patients by email to fill in web-based questionnaires. Data used for this present study were collected in the registry from October 2013 until February 2014. Our study procedures were approved by the Medical Ethical Committee of the VU Medical Center (2013/20).

Participants

MSK physicians were instructed to invite all consecutive patients who presented for the first time in MSK practice to participate. To evaluate cross-cultural validity, we used a part of the sample that was used in the original US calibration study. This sample consisted of 967 patients who were recruited through the website of the American Chronic Pain Association (ACPA), and who had at least one chronic pain condition for at least 3 months before participating in the survey.[18]

Measures

The PROMIS-PI item bank was developed as part of the NIH PROMIS project, and contains 40 items. The temporal context for all items is 7 days. Response categories are divided into three sets fitting the specific items: (1) not at all, a little bit, somewhat, quite a bit, very much, (2) never, rarely, sometimes, often, always, and (3) never, once a week or less, once every few days, once a day, every hour. The item bank was calibrated in a large US study on a population, including a community sample, and clinical samples of cancer patients, and of patients with chronic pain recruited through the American Chronic Pain Association (US-ACPA sample). Translation of the item bank into Dutch-Flemish was carried out by FACITtrans according to standard PROMIS methodology and approved by the PROMIS Statistical Center.[14] Our study population completed the full 40 item v1.1 Dutch-Flemish PROMIS Pain Interference item bank.

In addition to completing the PROMIS Pain Interference item bank, our study participants were asked to complete one of five condition-specific (legacy) instruments, according to their respective main complaint: the Roland Disability Questionnaire (RDQ),[19] the Neck Disability Index (NDI),[20] the Lower Extremity Function Scale (LEFS),[21] the Disabilities of the Arm, Shoulder and Hand (DASH),[22] and the Headache Impact Test (HIT-6),[23] for patients with low back pain, neck pain, lower extremity pain, upper extremity pain, or headache, respectively. The number of items and the range of scores are indicated in Table 4. All legacy instruments are frequently used in research and have been validated in Dutch populations.[24–30]

Statistical Analyses

Statistical analyses were carried out according to the PROMIS plan for psychometric evaluation and calibration of health-related quality of life item banks.[31] Descriptive analyses were carried out using SPSS 22, IBM, Armonk, NY, USA.

We evaluated dimensionality by confirmatory factor analyses (CFAs) and by assessing local independency in a

Graded Response Model. Model fit was evaluated by the following indices: the Comparative Fit Index (CFI, >0.95 for good fit), Tucker-Lewis Index (TLI, >0.95 for good fit), and the Root Mean Square Error of Approximation (RMSEA, <0.06 for good fit).[32] To evaluate the influence of multidimensionality, a bifactor model was fitted, and omega-H and explained common variance (ECV) were calculated. A high coefficient omega (> 0.80)[33] and a high ECV (> 0.60)[34] indicate that the risk of biased parameters when fitting multidimensional data into a unidimensional model is low. CFA was carried out with the R-package Lavaan (version 0.5–23.1097).

We assessed monotonicity as a measure of scalability with the R-package Mokken.[35] Mokken H was interpreted according to the following rules of thumb: unscalable if $H_{(i)} < 0.3$, weak if $0.3 \leq H_{(i)} < 0.4$, moderate if $0.4 \leq H_{(i)} < 0.5$, and strong if $H_{(i)} \geq 0.5$.[35,36]

An IRT model was used to study item fit and to calculate slope and threshold parameters. Items were considered to misfit if the *P* value is < 0.001. *T*-scores were calculated on the basis of US calibration parameters with the expected a priori method, using the R-package Mirt (version 1.24).[37] A *T*-score of 50 represents the mean score of the general population, with a SD of 10.

DIF was assessed for several age groups, for gender, and for language. We evaluated DIF for language (English *vs*. Dutch) by comparing our data with the data available from the US-ACPA sample used by Amtmann *et al* (N = 967).[38] DIF was analyzed using ordinal logistic regression models with the R-package Lordif (version 0.3–3),[39,40] with theta as an estimation of the trait level. The change in McFadden R^2 was used as an indicator of DIF, with a value of >0.02 serving as the critical value for rejecting the null hypothesis of no DIF.[39]

Construct validity was studied by testing hypotheses about the correlation of *T*-scores with the scores on several legacy instruments using SPSS statistics, version 22. Our hypothesis was that for the condition-specific subgroups of patients, the *T*-scores would correlate with the corresponding functional legacy instruments ($R > 0.50$).

TABLE 1. Demographics of Patient Sample: Age, Gender, Duration of Main Complaints, Primary Complaints, *T*-scores, and Scores on Legacy Instruments; Comparison With the US-ACPA Sample

	MSK Sample (N = 1677)	US-ACPA Sample (N = 967)
General background		
Age, mean (SD)	47 (14)	48 (11)
Gender (% female)	59	81
Duration of complaints number (%)		
< 3 mo	259 (15.9)	
3 mo–1 yr	330 (20.2)	53 (6)
> 1 yr	1041 (63.9)	876 (94)
Type of complaints number (%)*		
Low back pain	849 (50.6)	533 (55)
Neck or shoulder pain	349 (20.8)	447 (46)
Other back pain	134 (8.0)	
Lower extremity	163 (9.7)	
Headache	56 (3.3)	290 (22)
Upper extremity	37 (2.2)	
Other	85 (5.4)	
Pain interference scores		
T-score mean (SD)	58.1 (6.7)	68.6 (4.9)
T-score range	37.4 – 76.1	53.0 – 90.0
Legacy scores † mean (range)		
RDQ (N = 827)	8.9 (0–23)	
NDI (N = 269)	13.1 (0–33)	
LEFS (N = 159)	55.0 (11–80)	
DASH (N = 102)	31.6 (2.5–69.2)	
HIT-6 (N = 54)	60.2 (36–73)	

**In our study, only the main complaint could be scored, while in the US-ACPA study, multiple complaints could be indicated.*

†Roland Disability Questionnaire (RDQ); 24 items, range 0–24; Neck Disability Index (NDI); 10 items, range 0–50; Lower Extremity Function Scale (LEFS); 20 items, range 0–80; Disabilities of the Arm, Shoulder, and Hand (DASH); 30 items, range 0–100; Headache Impact Test-6 (HIT-6); six items, range 36–78.

RESULTS

Demographic Characteristics

Two thousand six hundred ten patients were asked to participate in our study; 2171 consented. Of these 2171 patients, 1745 (67%) answered the questionnaires. Because only the year of birth was reported, we excluded patients who could have been under the age of 18 at inclusion. A small number of patients failed to answer any item at all. After removal of patients under 18, and patients who had failed to complete the whole item bank, a sample of 1677 patients (64%) remained. Another 27 patients had a number of missing items. Model analyses were conducted on the sample of 1677 patients. *T* scores were calculated for the 1650 patients who had answered all items. Demographic data of our sample are presented in Table 1, together with the demographic data of the US-ACPA sample used in the DIF analyses. Half of the patients presented with a primary complaint of low back pain (50.6%), with or without sciatica, followed by neck or shoulder pain (20.8%).

Dimensionality

The fit of a one-factor model in the CFA resulted in a CFI of 0.903 (unscaled 0.978), a TLI of 0.897 (unscaled 0.978), and a RMSEA of 0.145 (unscaled 0.185). CFA fit indices indicated suboptimal fit of a one-factor model. Evaluation of the residual correlation matrix showed local dependence for 99 of the possible 780 ($1/2 \times 40 \times 39$) item pairs (12%), with residual correlations greater than 0.2.

The bifactor model contained one general factor and five group factors. The group factor items are presented in

TABLE 2. Group Factors From the Bifactor Analyses

Factor	Item Code*	Item
1	PAININ40	How often did pain prevent you from walking more than 1 mile?
	PAININ42	How often did pain prevent you from standing for more than 1 h?
	PAININ47	How often did pain prevent you from standing for more than 30 min?
2	PAININ50	How often did pain prevent you from sitting for more than 30 min?
	PAININ51	How often did pain prevent you from sitting for more than 10 min?
	PAININ54	How often did pain keep you from getting into a standing position?
	PAININ55	How often did pain prevent you from sitting for more than 1 h?
3	PAININ11	How often did you feel emotionally tense because of your pain?
	PAININ16	How often did pain make you feel depressed?
	PAININ24	How often was pain distressing to you?
	PAININ29	How often was your pain so severe you could think of nothing else?
	PAININ32	How often did pain make you feel discouraged?
	PAININ37	How often did pain make you feel anxious?
4	PAININ1	How difficult was it for you to take in new information because of pain?
	PAININ8	How much did pain interfere with your ability to concentrate?
	PAININ49	How much did pain interfere with your ability to remember things?
	PAININ56	How irritable did you feel because of pain?
5	PAININ9	How much did pain interfere with your day to day activities?
	PAININ18	How much did pain interfere with your ability to work (include work at home)?
	PAININ22	How much did pain interfere with work around the home?
	PAININ34	How much did pain interfere with your household chores?
	PAININ48	How much did pain interfere with your ability to do household chores?

PAININ indicates Pain Interference.

Table 2. For the bifactor model, the fit indices were higher than for the one-factor model. CFI was 0.964 (unscaled 0.996), the TLI was 0.961 (unscaled 0.996), and the RMSEA was 0.089 (unscaled 0.083). Omega-H was 0.97, and ECV was 0.81. In the bifactor model, no item pairs showed residual correlations greater than 0.2.

Monotonicity

Scalability coefficients are summarized in Table 3. All items had a scalability coefficient higher than the required 0.3, ranging from 0.367 (PAININ54; "How often did pain keep you from getting into a standing position?") to 0.686 (PAININ46; "How often did pain make it difficult for you to plan social activities?"). The scalability of the whole scale was $H = 0.596$, which is strong according to Mokken rules of thumb.

Item Fit, Item Parameters, and *T*-scores

After fitting an IRT model to our data, we studied item fit and the range of thetas covered. Table 3 summarizes the fit statistics of all items, and the slope and threshold parameters. There were two items with a S-X^2 below the threshold of 0.001 (PAININ31; "How much did pain interfere with your ability to participate in social activities?" and PAININ48; "How much did pain interfere with your ability to do household chores?"). Slope parameters ranged from 1.00 to 4.27, and threshold parameters ranged from −1.77 to 3.66. The average *T*-score of our study population was 58.1 (range 37.4–76.1, SD 6.7).

Differential Item Functioning

None of the items showed DIF for any of the age groups or for gender. Uniform DIF for language was demonstrated for one item: PAININ24 ("How often was pain distressing to you?") showed lower threshold parameters for the Dutch population. The influence of this item is depicted in Figure 1. It shows that, in theory, for patients with a similar trait level, there would be a difference of less than 0.6 points in expected score when using this DIF item only. However, this difference was negligible when using the item bank as a whole.

Construct Validity

Table 4 summarizes that the *T*-scores correlated highly (all $R > 0.50$) with the scores of the legacy instruments.

DISCUSSION

We studied the validity of the Dutch-Flemish version of the PROMIS Pain Interference item bank in a large population of patients presenting with predominant complaints of MSK pain. The DF-PROMIS-PI item bank showed suboptimal fit to a one-factor model in CFA and some local dependence. None of the items violated the monotonicity assumption. A bifactor model showed good fit, a high coefficient omega-H and ECV, and no local dependence. The item bank showed good IRT item fit, good coverage of the pain interference construct, and good construct validity.

CFA fit indices and the presence of local dependence suggested suboptimal unidimensionality. In a previous study validating the Dutch-Flemish PROMIS Pain Interference item bank in an outpatient rehabilitation population with chronic pain, Crins *et al*[15] reported a better fit (CFI 0.986, TLI 0.986, RMSEA 0.159). In the study of Crins *et al*,[15] however, unscaled indices were reported, where it is now thought that, due to non-normality of the data, scaled indices should be used. The unscaled indices in our study would suggest better evidence of unidimensionality as well (CFI 0.978, TLI 0.978, RMSEA 0.185). The US calibration

TABLE 3. Scalability, GRM Item Parameters, and Fit Statistics

	Item	Mokken	Slope	Category Threshold				Item Statistics*	
		H_i	a	B1	B2	B3	B4	S-X2	Prob X2
PAININ1	How difficult was it for you to take in new information because of pain?	0.574	2.21	−0.07	0.71	1.61	3.00	261.46	0.3276
PAININ3	How much did pain interfere with you enjoyment of life?	0.655	2.98	−0.88	−0.01	0.76	1.95	299.87	0.0052
PAININ5	How much did pain interfere with your ability to participate in leisure activities?	0.623	2.59	−1.07	−0.17	0.47	1.67	343.81	0.0016
PAININ6	How much did pain interfere with your close personal relationships?	0.650	3.38	−0.13	0.53	1.29	2.51	233.87	0.0887
PAININ8	How much did pain interfere with your ability to concentrate?	0.585	2.15	−0.62	0.19	1.07	2.28	344.66	0.0110
PAININ9	How much did pain interfere with your day to day activities?	0.641	2.67	−1.43	−0.27	0.59	1.91	263.09	0.1568
PAININ10	How much did pain interfere with your enjoyment of recreational activities?	0.635	2.78	−1.12	−0.13	0.50	1.64	292.91	0.0612
PAININ11	How often did you feel emotionally tense because of your pain?	0.588	2.11	−0.72	0.06	1.19	3.07	294.79	0.0934
PAININ12	How much did pain interfere with the things you usually do for fun?	0.637	2.83	−1.13	−0.12	0.50	1.63	319.29	0.0034
PAININ13	How much did pain interfere with your family life?	0.632	2.79	−0.55	0.23	1.06	2.23	263.00	0.1924
PAININ14	How much did pain interfere with doing your tasks away from home (*e.g.*, getting groceries, running errands)?	0.642	3.13	−0.16	0.52	1.15	2.06	247.34	0.3088
PAININ16	How often did pain make you feel depressed?	0.545	1.91	−0.18	0.65	1.87	3.64	273.65	0.3128
PAININ17	How much did pain interfere with your relationships with other people?	0.637	3.13	−0.16	0.52	1.40	2.50	268.72	0.0066
PAININ18	How much did pain interfere with your ability to work (include work at home)?	0.644	2.88	−0.84	−0.02	0.67	1.76	238.53	0.7197
PAININ19	How much did pain make it difficult to fall asleep?	0.452	1.30	−0.84	0.28	1.23	2.64	369.47	0.3677
PAININ20	How much did pain feel like a burden to you?	0.637	2.43	−1.77	−0.64	0.12	1.55	241.42	0.7049
PAININ22	How much did pain interfere with work around the home?	0.649	2.79	−1.32	−0.31	0.45	1.68	280.03	0.1084
PAININ24	How often was pain distressing to you?	0.507	1.50	−1.11	−0.21	1.36	3.26	366.90	0.0086
PAININ26	How often did pain keep you from socializing with others?	0.672	3.56	−0.46	0.21	1.11	2.34	206.22	0.3477
PAININ29	How often was your pain so severe you could think of nothing else?	0.598	2.34	−0.22	0.56	1.63	3.28	269.34	0.0937
PAININ31	How much did pain interfere with your ability to participate in social activities?	0.685	4.16	−0.44	0.23	0.89	1.87	307.30	0.0001
PAININ32	How often did pain make you feel discouraged?	0.613	2.34	−0.69	0.07	1.14	2.76	271.46	0.2028
PAININ34	How much did pain interfere with your household chores?	0.645	2.73	−1.24	−0.28	0.49	1.76	263.77	0.2923
PAININ35	How much did pain interfere with your ability to make trips from home that kept you gone for more than 2 h?	0.658	3.52	0.07	0.60	1.17	1.89	274.31	0.0122
PAININ36	How much did pain interfere with your enjoyment of social activities?	0.662	3.23	−0.74	0.01	0.67	1.75	246.63	0.3039
PAININ37	How often did pain make you feel anxious?	0.520	1.62	−0.53	0.34	1.74	3.54	333.69	0.0509

TABLE 3 (*Continued*)

	Item	Mokken	Slope	Category Threshold				Item Statistics*	
		H_i	a	B1	B2	B3	B4	S-X2	Prob X2
PAININ38	How often did you avoid social activities because it might make you hurt more?	0.639	3.13	−0.10	0.50	1.22	2.32	239.70	0.2535
PAININ40	How often did pain prevent you from walking more than 1 mile?	0.531	1.83	−0.21	0.34	0.96	1.84	384.26	0.0490
PAININ42	How often did pain prevent you from standing for more than one hour?	0.511	1.53	−0.79	−0.22	0.62	1.82	396.04	0.0925
PAININ46	How often did pain make it difficult for you to plan social activities?	0.686	4.27	−0.25	0.39	1.12	2.16	176.08	0.5898
PAININ47	How often did pain prevent you from standing for more than 30 min?	0.517	1.62	−0.39	0.20	1.08	2.13	364.26	0.3419
PAININ48	How much did pain interfere with your ability to do household chores?	0.659	3.05	−0.91	0.09	0.73	1.88	308.91	0.0009
PAININ49	How much did pain interfere with your ability to remember things?	0.562	2.10	0.29	0.96	1.82	2.94	296.11	0.0294
PAININ50	How often did pain prevent you from sitting for more than 30 min?	0.512	1.63	−0.13	0.58	1.47	2.71	303.43	0.7259
PAININ51	How often did pain prevent you from sitting for more than 10 min?	0.504	1.66	0.44	1.27	2.27	3.66	264.51	0.2667
PAININ52	How often was it hard to plan social activities because you did not know if you would be in pain?	0.644	3.29	0.06	0.62	1.29	2.16	239.88	0.2509
PAININ53	How often did pain restrict your social life to your home?	0.655	3.39	−0.03	0.58	1.36	2.61	206.16	0.4443
PAININ54	How often did pain keep you from getting into a standing position?	0.367	1.00	0.99	1.73	2.32	3.18	310.84	0.3505
PAININ55	How often did pain prevent you from sitting for more than 1 h?	0.487	1.48	−0.17	0.48	1.33	2.50	366.36	0.3970
PAININ56	How irritable did you feel because of pain?	0.571	2.03	−0.97	0.16	1.11	2.48	294.74	0.3330

*Statistical significance indicates poor item fit.

study reported good fit (CFI 0.974, TLI 0.997, RMSEA 0.175),[38] but did not state whether scaled or unscaled indices were reported. A secondary analysis on part of the US calibration sample reported suboptimal fit as well (CFI 0.90, TLI 0.90, RMSEA 0.135).[41] A cross-cultural validation study in a Spanish speaking population showed good fit (CFI 0.97, TLI 0.97, RMSEA 0.10) with no local dependency, without reporting whether scaled or unscaled indices were used.[42] Some authors have mentioned that unidimensionality could be hard to achieve when developing item banks for clinical measurement,[11,43] and it has been suggested that fit indices should not be regarded as measures of usefulness of a model.[44,45] In our study, the bifactor model, however, showed good fit (CFI 0.964, TLI 0.961, RMSEA 0.089), and the Omega-H coefficient and the ECV were high (0.97 and 0.81, respectively), indicating a low risk of biased parameters when treating the item bank as unidimensional.[33,34]

Item slope parameters ranged from 1.00 to 4.27 and item threshold parameters ranged from −1.77 to 3.66. Considering that under a normal distribution, 99.99% of the thetas will be in the range of −4 to +4, this range of threshold parameters represented a good coverage for a population of patients with pain. The item with the lowest slope parameter and both items with the lowest and the highest threshold parameters were the same as those reported by Crins *et al*.[15] In our study, two items showed poor fit, as opposed to one different item with poor fit reported by Crins *et al*,[15] and again one different item reported by Paz *et al*.[42]

DIF analyses did not show any DIF for age or gender; however, one item with DIF for language was found. The influence of this DIF for language on theta-scores was very limited. Crins *et al*[15] reported DIF for language for the same item (PAININ24) and also for item PAININ32, with a minimal impact on the Test Characteristics Curve as well. As the influence of DIF for language is very limited, we suggest that these items can be retained.

The strong correlations with several legacy instruments support construct validity of the DF-PROMIS-PI item bank. It is interesting to note that the five legacy instruments were developed to measure functional limitations for specific conditions. The correlation of one single item bank with several condition specific legacy instruments supports the generic use of the Pain Interference item bank. The good fit of the bi-factor model, together with the high omega-H and the high ECV, indicates that the PROMIS Pain Interference

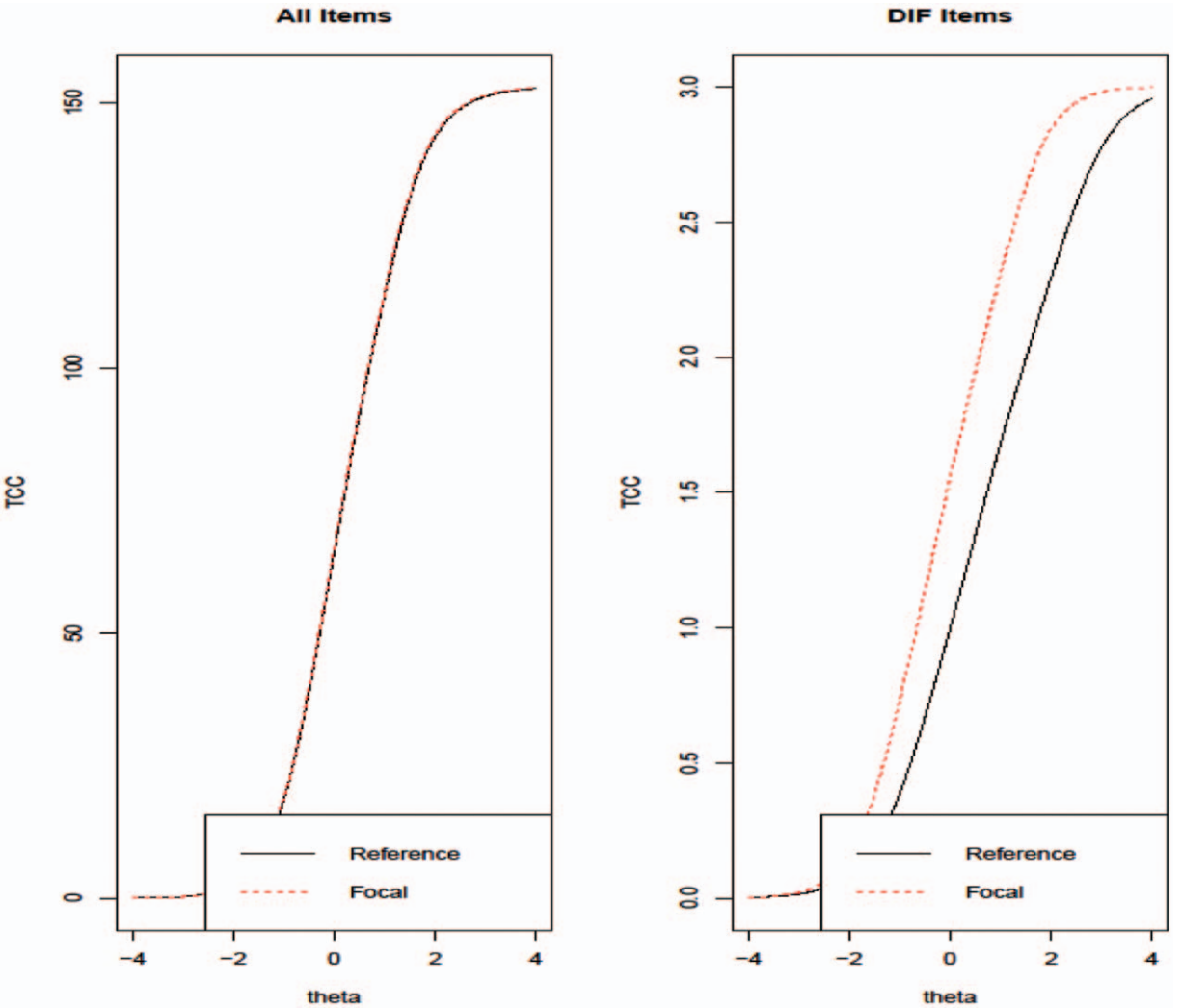

Figure 1. Test characteristics curves showing the influence of the DIF for language on theta estimates. The figure on the left shows the impact of DIF for language on theta scores when using the item bank as a whole. The figure on the right shows the impact when using the DIF item only. Reference line represents DF-PROMIS theta-scores.

item bank could be considered essentially unidimensional. The limited influence of DIF for language and the strong correlations with legacy instruments supports the validity of the Dutch-Flemish translation. Because of these properties, the PROMIS Pain Interference item bank can be considered suitable for use in both clinical research and practice, and can be used as a basis for short forms and computer adaptive testing.

TABLE 4. Mean Scores and Ranges of Legacy Instruments and Correlations Between PROMIS Pain Interference *T*-scores and Legacy Instruments

Instrument*			Measurement				Correlation	
	Items	Range	N	Mean	Min	Max	Expected *R*	Observed *R*
RDQ	24	0–24	827	8.9	0.0	23.0	>0.50	0.700
NDI	10	0–50	269	13.1	0.0	33.0	>0.50	0.687
LEFS	20	0–100	159	55.0	11.0	80.0	<−0.50	−0.754
DASH	30	0–80	102	31.6	2.5	69.2	>0.50	0.731
HIT-6	6	36–78	54	60.2	36.0	73.0	>0.50	0.527

Correlation with the LEFS is negative because higher disability is depicted in lower scores.

**DASH indicates Disabilities of the Arm Shoulder and Hand; HIT-6, Headache Impact Test; LEFS, Lower Extremity Function Scale; NDI, Neck Disability Index; RDQ, Roland Disability Questionnaire.*

CONCLUSION

The Dutch-Flemish v1.1 PROMIS Pain Interference item bank showed good IRT item fit, good coverage of the pain interference trait, and good construct validity. None of the items showed DIF for age or gender. One item showed minimal DIF for language. CFA and analyses of local independence showed evidence of multidimensionality, but omega-H and ECV were high, indicating a low risk of biased parameters when assuming unidimensionality. We conclude that our results support the validity of the DF-PROMIS-Pain Interference item bank, and that the item bank can be used as a basis for short forms and computer adaptive testing in clinical research and in clinical practice.

➢ Key Points

- ❑ New PROMS based upon Item Response Theory offer advantages compared with classical PROMS.
- ❑ These new PROMS consist of item banks that have to be validated in various populations.
- ❑ We studied the validity of the Dutch-Flemish PROMIS Pain Interference item bank in a large population of patients with musculoskeletal complaints.

Acknowledgments

We would like to thank all members of the Dutch Association for Musculoskeletal Medicine who cooperated in this study. We also would like to thank K. Uegaki for reviewing the manuscript.

References

1. Cella D, Gershon R, Lai JS, et al. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res* 2007;16 (suppl 1):133–41.
2. Cella D, Riley W, Stone A, 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.
3. de Vet HC, Terwee CB, Mokkink LB, et al. *Measurement in Medicine*. Cambridge: Cambridge University Press; 2011.
4. 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.
5. Fries JF, Krishnan E, Rose M, et al. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Res Ther* 2011; 13:R147.
6. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000;38 (9 suppl):II28–42.
7. Tugwell P, Knottnerus JA, Idzerda L. Tailoring patient reported outcome measurement. *J Clin Epidemiol* 2010;63:1165–6.
8. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. *J Rheumatol* 2007;34:1426–31.
9. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Qual Life Res* 2007;16 (Suppl 1):187–94.
10. Haley SM, Ni P, Hambleton RK, et al. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. *J Clin Epidemiol* 2006;59:1174–82.
11. Reise SP, Waller NG. Item response theory and clinical measurement. *Annu Rev Clin Psychol* 2009;5:27–48.
12. Fries JF, Krishnan E. What constitutes progress in assessing patient outcomes?. *J Clin Epidemiol* 2009;62:779–80.
13. Deyo RA, Dworkin SF, Amtmann D, et al. Focus article: report of the NIH task force on research standards for chronic low back pain. *Eur Spine J* 2014;23:2028–45.
14. Terwee CB, Roorda LD, de Vet HC, 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.
15. Crins MH, Roorda LD, Smits N, et al. Calibration and validation of the Dutch-Flemish PROMIS pain interference item bank in patients with chronic pain. *PLoS One* 2015;10:e0134094.
16. Schuller W, Ostelo R, Rohrich DC, et al. Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients. *BMC Musculoskelet Disord* 2017;18:512.
17. Lamberts HWM. *International Classification of Primary Care (ICPC)*. Oxford: Oxford University Press; 1987.
18. Revicki DA, Chen WH, Harnam N, et al. Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain* 2009;146:158–69.
19. 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.
20. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. *J Manipulative Physiol Ther* 1991;14:409–15.
21. Binkley JM, Stratford PW, Lott SA, et al. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. *Phys Ther* 1999;79:371–83.
22. 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.
23. Kosinski M, Bayliss MS, Bjorner JB, et al. A six-item short-form survey for measuring headache impact: the HIT-6. *Qual Life Res* 2003;12:963–74.
24. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain* 1996;65:71–6.
25. Beurskens AJ, de Vet HC, Koke AJ, et al. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. *Spine (Phila Pa 1976)* 1995;20:1017–28.
26. Hoogeboom TJ, de Bie RA, den Broeder AA, et al. The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study. *BMC Musculoskelet Disord* 2012;13:117.
27. Jorritsma W, de Vries GE, Geertzen JH, et al. Neck Pain and Disability Scale and the Neck Disability Index: reproducibility of the Dutch Language Versions. *Eur Spine J* 2010;19:1695–701.
28. Jorritsma W, Dijkstra PU, de Vries GE, et al. Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. *Eur Spine J* 2012;21:2550–7.
29. Martin M, Blaisdell B, Kwong JW, et al. The Short-Form Headache Impact Test (HIT-6) was psychometrically equivalent in nine languages. *J Clin Epidemiol* 2004;57:1271–8.
30. Veehof MM, Sleegers EJ, van Veldhoven NH, et al. 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.
31. Reeve BB, Hays RD, Bjorner JB, 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 (5 suppl 1):S22–31.

32. Hu LT, Bentler P. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equat Model* 1999;1999:1–55.
33. Rodriguez A, Reise SP, Haviland MG. Applying bifactor statistical indices in the evaluation of psychological measures. *J Pers Assess* 2016;98:223–37.
34. Reise SP, Scheines R, Widaman KF, et al. Multidimensionality and structural coefficient bias in structural equation modeling: a bifactor perspective. *Educ Psychol Meas* 2013;73:5–26.
35. van der Ark LA. Mokken scale analysis in R. *J Stat Softw* 2007; 1–19.
36. van der Ark LA, Croon MA, Sijtsma K. Mokken Scale analysis for dichotomous items using marginal models. *Psychometrika* 2008; 73:183–208.
37. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. *J Stat Softw* 2012;48:1–29.
38. Amtmann D, Cook KF, Jensen MP, et al. Development of a PROMIS item bank to measure pain interference. *Pain* 2010; 150:173–82.
39. 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.
40. Crane PK, Gibbons LE, Jolley L, et al. Differential item functioning analysis with ordinal logistic regression techniques. *Med Care* 2006;44:S115–23.
41. Kim J, Chung H, Amtmann D, et al. Measurement invariance of the PROMIS pain interference item bank across community and clinical samples. *Qual Life Res* 2013;22:501–7.
42. Paz SH, Spritzer KL, Reise SP, et al. Differential item functioning of the patient-reported outcomes information system (PROMIS((R))) pain interference item bank by language (Spanish versus English). *Qual Life Res* 2017;26:1451–62.
43. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Qual Life Res* 2009;18: 447–460.
44. Browne MW, Cudeck R. Alternative ways of assessing model fit. *Sociol Method Res* 1992;21:230–58.
45. Iacobucci D. Structural equations modeling: fit indices, sample size, and advanced topics. *J Consum Psychol* 2010; 20:90–8.