Quality of Life Research (2019) 28:1231–1243
https://doi.org/10.1007/s11136-018-2087-x

Differential item functioning of the PROMIS physical function, pain interference, and pain behavior item banks across patients with different musculoskeletal disorders and persons from the general population

Martine H. P. Crins[1] · Caroline B. Terwee[2] · Oguzhan Ogreden[2] · Wouter Schuller[2,3] · Paul Dekker[1,4] · Gerard Flens[5] · Daphne C. Rohrich[2] · Leo D. Roorda[1]

Accepted: 10 December 2018 / Published online: 2 January 2019
© Springer Nature Switzerland AG 2019

Abstract

Purpose To investigate the validity of comparisons across patients with different musculoskeletal disorders and persons from the general population by evaluating differential item functioning (DIF) for the PROMIS physical function (PROMIS-PF), pain interference (PROMIS-PI), and pain behavior (PROMIS-PB) item banks.

Methods Patients with chronic pain, rheumatoid arthritis (RA), or osteoarthritis (OA); patients receiving physiotherapy (PT); and persons from the Dutch general population completed the full Dutch-Flemish PROMIS-PF (121-items), PROMIS-PI (40-items), or PROMIS-PB (39-items) banks. DIF was assessed with ordinal logistic regression models and McFadden's pseudo R^2-change of $\geq 2\%$ as critical value. The impact of DIF on item scores and the *T*-scores per bank was examined by inspecting item characteristic curves (ICCs) and test characteristic curves (TCCs).

Results 2762 patients with chronic pain, 2029 with RA, 1247 with OA, 805 receiving PT, and 1310 healthy persons participated. For the PROMIS-PF, 25 out of 121 items were flagged for DIF, of which 10 items were flagged in multiple comparisons. For the PROMIS-PI, only 2 out of 40 items were flagged for DIF and for the PROMIS-PB, only 3 out of 39 items. Most DIF items had R^2 values just above the critical value of 2% and all showed uniform DIF. The ICCs and TCCs showed that the magnitude and impact of DIF on the item and *T*-scores were negligible.

Conclusions This study supports the universal applicability of PROMIS across (patient) populations. Comparisons across patients with different musculoskeletal disorders and persons from the general population are valid, when applying the PROMIS-PF, PROMIS-PI, and PROMIS-PB banks.

Keywords PROMIS · Chronic pain · Rheumatoid arthritis · Hip or knee osteoarthritis · General population · Differential item functioning

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-018-2087-x) contains supplementary material, which is available to authorized users.

✉ Martine H. P. Crins
m.crins@reade.nl
https://www.reade.nl

[1] Amsterdam Rehabilitation Research Center | Reade, Dr. Jan van Breemenstraat, Postbus 58271, 1040 HG Amsterdam, The Netherlands

[2] Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit Amsterdam, Boelelaan 1117, Amsterdam, The Netherlands

[3] Spineclinic, Zaandam, The Netherlands

[4] Department of Rehabilitation, Leiden University Medical Center, Leiden, The Netherlands

[5] Foundation for Benchmarking Mental Health Care, Bilthoven, The Netherlands

Introduction

The Patient-Reported Outcomes Measurement Information System (PROMIS) is a set of instruments measuring patient-reported health [1, 2]. PROMIS instruments consist of item banks, a set of items (questions) that measure one health domain. These banks can be applied as short forms (fixed length subsets of items out of the item bank) or highly efficient computerized adaptive tests (CAT). A CAT is a computer-administered measure in which successive items are selected by a computer algorithm informed by the responses to previous items. Persons generally have to complete only a small number (3–7) of highly informative and relevant items to obtain a reliable ($r = 0.90$) score. Overall, PROMIS instruments are less burdensome, have less measurement error, have better content validity than traditional Patient-Reported Outcome Measures, and are easy to interpret [3–6].

One of the goals when developing PROMIS was to create measures that would be universally applicable. A universal measure should be applicable within multiple (patient) populations and should also be valid for comparisons across (patient) populations. PROMIS item banks have been developed using item response theory (IRT). Validity of comparisons between populations, in the context of IRT, is plausible if the item parameters are equivalent between the comparison populations at issue. Equivalence of item parameters implies the absence of the so-called differential item functioning (DIF) [7–9].

Three of the most commonly used PROMIS item banks are the PROMIS physical function (PROMIS-PF), pain interference (DF-PROMIS-PI), and pain behavior (PROMIS-PB) banks. Those banks showed good psychometric properties for cross-sectional use within different (patient) populations [10–17]. Furthermore, some studies on the PROMIS-PF, PROMIS-PI, and PROMIS-PB banks evaluated DIF with respect to language (Dutch-Flemish vs. English, Spanish vs. English, German vs. English) and demographic variables, such as age and gender. In these studies, either no DIF was found or the observed DIF had a negligible impact on the *T*-scores [10–23]. To our knowledge, no studies so far have examined DIF across patient populations for the PROMIS-PF, PROMIS-PI, and PROMIS-PB banks.

In patients with musculoskeletal disorders, physical functioning and pain are core outcomes. Health care providers, including rehabilitation physicians, rheumatologists, orthopedic surgeons, and physiotherapists, provide care to patients with different musculoskeletal disorders [e.g., patients with chronic pain, rheumatoid arthritis (RA), or osteoarthritis (OA)]. It would be beneficial to all these providers if one measure can be used in all these patient populations, in patients who have multiple of these disorders, and also to compare these populations with each other and with healthy persons. Therefore, the aim of the study is to investigate the validity of comparisons across patients with different musculoskeletal disorders and persons from the general population by evaluating DIF for the Dutch-Flemish PROMIS-PF, PROMIS-PI, and PROMIS-PB banks.

Methods

Samples

We used five datasets to study DIF across patient populations for the PROMIS-PF (V1.2), PROMIS-PI (V1.1), and PROMIS-PB (V1.1) banks. All datasets contained cross-sectional data including multiple item banks and most datasets combined response data of more than one sample.

The first dataset consisted of Dutch patients with chronic musculoskeletal pain (PAIN dataset). We used, firstly, data of a sample of patients with chronic pain from the Amsterdam Pain (AMS-PAIN) cohort. These data were collected at the rehabilitation outpatient department of Reade, a care center for rehabilitation and rheumatology, in the Netherlands (PROMIS-PF, $n = 1247$ [16]; PROMIS-PI, $n = 1085$ [14]; and, PROMIS-PB, $n = 1042$ [13]). We used, secondly, data of Dutch patients with chronic pain registered at practices of 31 participating physicians specialized in musculoskeletal medicine in the Netherlands (PROMIS-PI, $n = 1677$ [21]; PROMIS-PB, $n = 1602$ [20]). So, with respect to the PROMIS-PF bank, the dataset consisted of patients from the AMS-PAIN cohort only (AMS-PAIN dataset), whereas, with respect to the PROMIS-PI and PROMIS-PB banks, the dataset consisted of the two combined chronic pain samples (PAIN dataset). A preliminary analysis indicated no DIF between these two chronic pain samples for the PROMIS-PI and PROMIS-PB banks, supporting our decision to combine these two samples.

The second dataset comprised Dutch and Flemish patients with RA (RA dataset). The Dutch sample consisted of patients with RA from the Amsterdam Rheumatoid Arthritis cohort and the data were collected at the rheumatology outpatient department of Reade (PROMIS-PI, $n = 1370$; PROMIS-PB, $n = 1005$) [19]. The Flemish sample consisted of patients with RA from an arthritis cohort from University Hospitals Leuven, Flanders, the Dutch speaking part of Belgium (PROMIS-PI, $n = 682$; PROMIS-PB, $n = 549$) [19]. In a previous study, we found no DIF for language (Dutch vs. Flemish) for these item banks [19], which legitimizes the merging of the data from the two samples.

The third dataset consisted of Dutch patients with hip or knee OA (OA dataset). We used, firstly, response data of a sample of patients with hip or knee OA from the Amsterdam Osteoarthritis (AMS-OA) cohort. These data were collected at the rehabilitation outpatient department of Reade (PROMIS-PF, $n = 425$; PROMIS-PI, $n = 425$ [unpublished]). We used, secondly, response data of patients with early hip or knee OA from the Cohort Hip and Cohort Knee (CHECK) cohort [24]. These data were collected during a 10-year follow-up measurement at Erasmus Medical Center Rotterdam, Kennemer Gasthuis Haarlem, Leiden University Medical Center, Maastricht University Medical Center, Martini Hospital Groningen/Allied Health Care Center for Rheumatology and Rehabilitation Groningen, Medical Spectrum Twente Enschede/Ziekenhuisgroep Twente, Reade, Center for Rehabilitation and Rheumatology, St Maartenskliniek Nijmegen, University Medical Center Utrecht, and Wilhelmina Hospital Assen (PROMIS-PF, $n = 822$ [25]). So, with respect to the PROMIS-PF bank, the dataset consisted of the two combined datasets (OA dataset). A preliminary analysis indicated one item with DIF between these two OA samples for the PROMIS-PF bank, but the impact of this DIF was negligible, supporting our decision to combine these two samples. With respect to the PROMIS-PI bank, the dataset consisted of patients from the AMS-OA cohort only (AMS-OA dataset).

The fourth dataset consisted of Dutch patients who received any kind of physiotherapy (PT) in primary care in the year prior to completing the questionnaire (PT dataset, PROMIS-PF, $n = 805$ [17]). Forty-nine percent of the patients consulted PT because of disorders of muscles, bones or joints, and twelve percent as part of recovery after a surgery [17].

The fifth dataset represented a Dutch general population sample (GEN) (GEN dataset). Participants were recruited from an existing internet panel of the general Dutch panel, polled by a certified company (Desan Research Solutions) (PROMIS-PF, $n = 1310$; PROMIS-PI, $n = 1052$; and PROMIS-PB, $n = 745$ [unpublished]). The sample was representative for the Dutch general population (maximum of 2.5% deviation) with respect to distribution of age, gender, education, region, and ethnicity, according to data from Statistics Netherlands in 2016.

Measures

The participants completed a paper-and-pencil or web-based survey which included, among others, demographic and clinical characteristics, and the Dutch-Flemish versions of the full PROMIS-PF, PROMIS-PI, or PROMIS-PB [26] banks.

The PROMIS-PF bank assesses a wide range of activities, from self-care (activities of daily living) to more complex activities that require a combination of skills (i.e., strenuous activities such as playing tennis, bicycling or jogging). The PROMIS-PF bank (V1.2) consists of 121 items, including items about functioning of the axial regions (neck and back), the upper and lower extremities, and ability to carry out instrumental activities of daily living (i.e., housework, shopping) [10]. There is no time frame set for the items, but current status is inferred. There are three different 5-point Likert response scales. For the PROMIS-PF bank, higher *T*-scores indicate higher (i.e., better) levels of physical function. The PROMIS-PF bank showed good psychometric properties for cross-sectional use within different populations [10, 15–17].

The PROMIS-PI bank assesses 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 [27]. The PROMIS-PI bank (V1.1) consists of 40 items. The time frame is the past 7 days, and the bank uses three different 5-point Likert response scales [11, 27]. For the PROMIS-PI bank, higher *T*-scores indicate higher (i.e., worse) levels of pain interference. The PROMIS-PI bank showed good psychometric properties for cross-sectional use within different populations [11, 14, 19, 21].

The PROMIS-PB bank measures self-reported external manifestations of pain: behaviors that typically indicate to others that an individual is experiencing pain [28]. The PROMIS-PB bank (V1.1) contains 39 items. Patients rate how frequently they engaged in the pain behaviors in the past 7 days on a 6-point Likert response scale [12]. We excluded patients who endorsed the 'had no pain' response category on any of the items, resulting in IRT analyses with five response options [13, 29]. This is in line with later analyses of the PROMIS pain behavior item bank (resulting in version 2.0) where the researchers decided to develop version 2.0 only for patients with pain, and the response option "had no pain" is no longer used [29]. For the PROMIS-PB bank, higher *T*-scores indicate higher levels (i.e., worse) of pain behavior. The PROMIS-PB bank showed also good psychometric properties for cross-sectional use within different populations [12, 13, 19, 20].

PROMIS scores are expressed as *T*-scores, and a mean of 50 represents the average score of the general population with a standard deviation of 10.

Statistical analysis

In order to study the validity of comparisons across (patient) populations, we evaluated DIF across (patient) populations. For the PROMIS-PF bank, we made six comparisons: AMS-PAIN vs. OA, AMS-PAIN vs. PT, AMS-PAIN vs. GEN, OA vs. PT, OA vs. GEN, and PT vs. GEN (Table 1). With respect to the PROMIS-PI bank, we also made six comparisons: PAIN vs. RA, PAIN vs. AMS-OA, PAIN vs. GEN, RA vs. AMS-OA, RA vs. GEN, and AMS-OA vs. GEN (Table 1). For the PROMIS-PB bank, we

Table 1 Demographic and clinical characteristics of the patient populations with different musculoskeletal disorders and the general population sample per PROMIS item bank

	PROMIS-PF				PROMIS-PI				PROMIS-PB		
	AMS-Pain	OA	PT	GEN	PAIN	RA	AMS-OA	GEN	PAIN	RA	GEN
Number of cases	1247	1247	805	1310	2762	2029	425	1052	2644	1554	745
Age											
Mean (SD)	48 (13)	69 (8)	53 (14)	51 (17)	48 (13)	59 (13)	69 (9)	52 (16)	48 (13)	59 (12)	53 (16)
Range	18–85	25–92	18–88	19–87	18–91	19–94	25–92	19–88	18–91	19–94	19–88
Gender %											
Male	22	23	41	47	33	31	28	47	33	31	46
Female	78	77	59	53	67	69	72	53	67	69	54
Disease duration %											
1–6 months	2	0	30	NA	19	0.3	0	NA	19	0.2	NA
6–12 months	4	0.5	21	NA	9	2	0.5	NA	10	2	NA
1–2 years	13	3	18	NA	14	7	3	NA	14	6	NA
2–5 years	29	21	11	NA	21	16	21	NA	21	15	NA
>5 years	52	76	20	NA	37	75	76	NA	36	76	NA
T-scores											
Mean (SD)	35.7 (7.4)	42.9 (7.4)	48.1 (9.5)	49.8 (10.9)	60.4 (7.4)	53.6 (9.8)	56.0 (8.1)	54.8 (8.4)	58.4 (4.7)	56.7 (5.1)	55.1 (6.6)
Range	7.5–73.5	12.1–73.5	20.2–73.5	6.0–73.5	37.5–85.1	37.5–75.7	37.5–74.3	37.5–82.6	38.4–83.3	38.4–73.4	38.4–83.3

AMS-PAIN patients with chronic pain from the Amsterdam Pain cohort, *AMS-OA* patients with hip or knee OA from the Amsterdam Osteoarthritis cohort, *GEN* general population sample, *OA* patients with hip or knee osteoarthritis (two combined OA samples), *PAIN* patients with chronic musculoskeletal pain (two combined chronic pain samples), *PT* physiotherapy patients, *RA* patients with rheumatoid arthritis made three comparisons: PAIN vs. RA, PAIN vs. GEN, and RA vs. GEN (Table 1).

DIF analyses evaluate if persons from different populations (e.g., OA vs. GEN) with similar levels of the domain (e.g., physical function) respond similarly to the items [7–9]. The absence of DIF implies valid comparisons of *T*-scores between the populations at issue. There are two kinds of DIF: uniform and non-uniform [7–9]. Uniform DIF exists if the magnitude of DIF is consistent across the entire range of the domain. Non-uniform DIF exists if the magnitude of DIF varies across the domain.

DIF was evaluated with the R package Lordif (version 0.3-3), which uses an ordinal logistic regression framework [7, 30–32]. Three models were formed, of which Fig. 1 shows a simplified version of the model originally published by Choi et al. [31]. These models will be explained using the physical function domain as example. Model 1, the base model, assumes that the persons' level of physical function (theta or, in the context of PROMIS, the *T*-score) only predicts the persons' item response. Model 2 posits that, in addition to the level of physical function, the persons' item response is predicted by population membership (e.g., OA vs. GEN). Uniform DIF is identified if model 2 predicts the item response better than model 1. Model 3 includes an interaction term between the level of physical function and population membership and posits that the relation between the level of physical function and the persons' item response is different across the populations being compared. Non-uniform DIF is present if model 3 predicts the item response better than model 2.

There are several criteria for identifying DIF and to date PROMIS researchers mostly have used the criterion of R^2-change of ≥ 0.02 [13, 15, 18, 23, 30, 33, 34]. We used in this study McFadden's pseudo R^2-change between two models of 0.02 as the critical value to flag for possible DIF [35].

If items were flagged for DIF, the impact of DIF on the item score and the *T*-score were examined by plotting item characteristics plots and test characteristic curves (TCCs), respectively. The item characteristic plots include four plots:

1. The item characteristic curves (ICCs) or item true score functions per population. This plot illustrates which population has higher item scores across levels of theta.
2. The absolute difference between the ICCs or differences in item true score functions. This plot shows the difference in item scores between the populations across levels of theta.
3. The item response functions, including the item slope and threshold parameters, per population. This plot visualizes which population has higher probabilities of endorsing the response categories at issue across levels of theta. The thresholds indicate the level of theta necessary to respond above this threshold with 0.50 probability.
4. The impact weighted by density. This plot shows the absolute difference in item scores weighted by the theta distribution of the samples [30–32].

The TCCs show per item bank and per population comparison of the test score (raw summary score) for all items (ignoring DIF) in the left plot, and the test scores for only the items having DIF in the right plot [30–32]. The area between the two curves within each plot provides an indication of the impact of DIF on the test score.

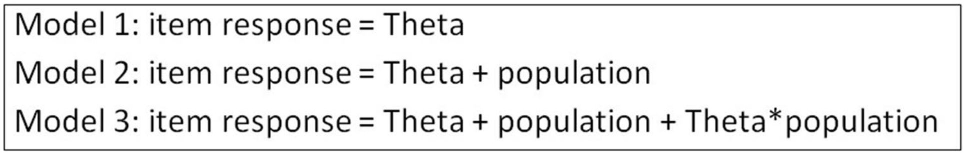
Results

Samples descriptives

Table 1 summarizes the demographic and clinical characteristics and PROMIS-PF, PROMIS-PI, and PROMIS-PB *T*-scores per dataset. The average age in the different samples, the proportion male vs. female, and the distribution of the duration of the conditions, match with the demographic and clinical characteristics in comparable populations [13, 14, 16, 17, 19–21]. Most clinical samples showed reduced physical function levels and elevated pain interference and pain behavior levels compared to the general population.

Differential item functioning

Table 2 summarizes the results. For the PROMIS-PF bank, 25 out of 121 items were flagged for DIF, of which 10 items were flagged for DIF in multiple comparisons and of which 3 items are present in the PROMIS-PF 20a short form. For the PROMIS-PI bank, only 2 out of 40 items were flagged for DIF. Both items are not present in any PROMIS-PI short form. For the PROMIS-PB bank, only 3 out of 39 items were flagged for DIF, of which all 3 items are present in the PROMIS 7a short form. All DIF items showed uniform


Model 1: item response = Theta

Model 2: item response = Theta + population

Model 3: item response = Theta + population + Theta*population

Fig. 1 Models used in the ordinal logistic regression

Table 2 Results of the DIF analysis of the PROMIS-PF, PROMIS-PI, and PROMIS-PB banks

PROMIS item banks	Populations compared	Items with DIF	McFadden R^2 values	Slope; and thresholds parameters
Physical function	Chronic pain (AMS-PAIN) vs. osteoarthritis (OA)	PFA19r1 (item 11)—"Are you able to run or jog for two miles (3 km)?"	$R^2_{12}=.027$ $R^2_{23}=.000$	**AMS-PAIN** = 1.70; 0.44, 0.97, 1.39, 2.10 OA = 1.74; 1.10, 1.69, 2.28, 3.11
		PFA31r1 (item 19)—"Are you able to get up from the floor from lying on your back without help?"	$R^2_{12}=.032$ $R^2_{23}=.001$	**AMS-PAIN** = 2.32; − 1.69, − 0.78, − 0.25, 0.50 OA = 2.21; − 1.12, − 0.26, 0.32, 1.27
		PFA39r1 (item 27)—"Are you able to run at a fast pace for two miles (3 km)?"	$R^2_{12}=.035$ $R^2_{23}=.000$	**AMS-PAIN** = 1.52; 0.59, 1.16, 1.64, 2.32 OA = 1.44; 1.56, 2.19, 2.71, 3.61
		PFA41 (item 29)—"Are you able to squat and get up?"	$R^2_{12}=.039$ $R^2_{23}=.001$	**AMS-PAIN** = 2.28; − 1.46, − 0.70, − 0.13, 0.72 OA = 1.70; − 0.60, − 0.01, 0.58, 1.67
		PFB13 (item 51)—"Are you able to carry a shopping bag or briefcase?"	$R^2_{12}=.037$ $R^2_{23}=.003$	AMS-PAIN = 2.65; − 1.66, − 0.78, − 0.25, 0.66 **OA** = 3.25; − 1.68, − 1.21, − 0.74, − 0.02
		PFB14 (item 52)—"Are you able to take a tub bath?"	$R^2_{12}=.023$ $R^2_{23}=.000$	**AMS-PAIN** = 2.79; − 1.93, − 1.43, − 0.91, − 0.26 OA = 2.18; − 1.15, − 0.79, − 0.47, 0.10
		PFC7r1 (item 79)—"Are you able to run five miles (8 km)?"	$R^2_{12}=.041$ $R^2_{23}=.001$	**AMS-PAIN** = 1.33; 1.04, 1.67, 2.24, 2.94 OA = 1.54; 1.96, 2.58, 3.09, 3.87
		PFC13r1 (item 80)—"Are you able to run 100 yards (100 m)?"	$R^2_{12}=.030$ $R^2_{23}=.000$	**AMS-PAIN** = 2.06; − 0.50, 0.10, 0.43, 1.11 OA = 1.96; 0.19, 0.68, 1.14, 1.98
		PFC33r1 (item 84)—"Are you able to run ten miles (16 km)?"	$R^2_{12}=.065$ $R^2_{23}=.007$	**AMS-PAIN** = 0.95; 2.01, 2.81, 3.37, 4.34 OA = 1.59; 2.78, 3.37, 3.63, 4.32
		PFC40 (item 87)—"Are you able to kneel on the floor?"	$R^2_{12}=.052$ $R^2_{23}=.002$	**AMS-PAIN** = 2.50; − 1.35, − 0.70, − 0.33, 0.33 OA = 1.95; − 0.59, 0.03, 0.48, 1.37
		PFC41 (item 88)—"Are you able to sit down in and stand up from a low, soft couch?"	$R^2_{12}=.036$ $R^2_{23}=.001$	**AMS-PAIN** = 2.37; − 2.66, − 0.94, − 0.40, 0.39 OA = 2.40; − 1.55, − 0.42, 0.20, 1.18
		PFC45r1 (item 89)[a]—"Are you able to sit on and get up from the toilet?"	$R^2_{12}=.024$ $R^2_{23}=.001$	**AMS-PAIN** = 2.97; − 2.87, − 1.64, − 1.06, − 0.47 OA = 2.76; − 2.60, − 1.36, − 0.79, 0,02
		PFB50 (item 99)—"How much difficulty do you have doing your daily physical activities, because of your health?"	$R^2_{12}=.029$ $R^2_{23}=.001$	AMS-PAIN = 2.80; − 0.57, 0.29, 1.88 **OA** = 3.58; − 1.05, − 0.12, 0.86
		PFC54 (item 104)—"Does your health now limit you in getting in and out of the bathtub?"	$R^2_{12}=.032$ $R^2_{23}=.001$	**AMS-PAIN** = 2.48; − 2.02, − 1.34, − 0.64, − 0.06 OA = 2.28; − 1.05, − 0.60, − 0.13, 0.45
	Chronic Pain (AMS-PAIN) vs. physiotherapy (PT)	PFA51 (item 38)[a]—"Are you able to sit on the edge of a bed?"	$R^2_{12}=.108$ $R^2_{23}=.000$	**AMS-PAIN** = 3.39; − 2.73, − 2.09, − 1.62, − 1.14 PT = 2.76; − 1.15, − 1.02, − 0.66, − 0.18
		PFB13 (item 51)—"Are you able to carry a shopping bag or a briefcase?"	$R^2_{12}=.021$ $R^2_{23}=.001$	AMS-PAIN = 3.09; − 1.62, − 0.87, − 0.41, 0.37 **PT** = 3.04; − 1.69, − 1.21, − 0.78, − 0.20

Table 2 (continued)

PROMIS item banks	Populations compared	Items with DIF	McFadden R^2 values	Slope; and thresholds parameters
	Chronic Pain (AMS-PAIN) vs. general population (GEN)	PFB13 (item 51)—"Are you able to carry a shopping bag or a briefcase?"	$R^2_{12}=.020$ $R^2_{23}=.001$	AMS-PAIN = 3.42; − 1.74, − 1.05, − 0.64, − 0.07 **GEN** = 3.46; − 1.83, − 1.38, − 0.96, − 0.43
		PFB49 (item 112)—"Does your health now limit you in going for a short walk (less than 15 min)?"	$R^2_{12}=.035$ $R^2_{23}=.000$	**AMS-PAIN** = 3.17; − 2.15, − 1.54, − 0.99 GEN = 3.16; − 1.56, − 1.05, − 0.51
	Osteoarthritis (OA) vs. physiotherapy (PT)	PFA51 (item 38)[a]—"Are you able to sit on the edge of a bed?"	$R^2_{12}=.220$ $R^2_{23}=.001$	**OA** = 2.90; − 2.57, − 1.99 PT = 2.38; − 1.19, − 0.63
		PFA56 (item 43)[a]—"Are you able to get in and out of a car?"	$R^2_{12}=.022$ $R^2_{23}=.001$	OA = 2.45; − 2.27, − 1.38, − 0.13 **PT** = 2.55; − 2.40, − 1.71, − 0.61
		PFB50 (item 99)—"How much difficulty do you have doing your daily physical activities, because of your health?"	$R^2_{12}=.023$ $R^2_{23}=.001$	**OA** = 3.67; − 1.76, − 0.83, 0.14 PT = 2.85; − 1.54, − 0.53, 0.51
	Osteoarthritis (OA) vs. general population (GEN)	PFA13 (item 6)—"Are you able to exercise for an hour?"	$R^2_{12}=.034$ $R^2_{23}=.005$	**OA** = 3.24; − 1.58, − 1.13, − 0.65, 0.05 GEN = 2.20; − 1.30, − 0.74, − 0.17, 0.67
		PFA41 (item 29)—"Are you able to squat and get up?"	$R^2_{12}=.029$ $R^2_{23}=.002$	OA = 2.12; − 1.27, − 0.80, − 0.32, 0.57 **GEN** = 2.65; − 1.59, − 1.17, − 0.82, − 0.12
		PFA50 (item37)—"Are you able to brush your teeth?"	$R^2_{12}=.023$ $R^2_{23}=.001$	**OA** = 3.29; − 2.88, − 2.56, − 2.02 GEN = 2.62; − 2.54, − 2.15, − 1.79
		PFA56 (item43)[a]—"Are you able to get in and out of a car?"	$R^2_{12}=.030$ $R^2_{23}=.005$	OA = 2.95; − 2.08, − 1.34, − 0.30 **GEN** = 3.71; − 2.06, − 1.56, − 0.79
		PFC33r1 (item 84)—"Are you able to run ten miles (16 km)?"	$R^2_{12}=.022$ $R^2_{23}=.004$	OA = 2.02; 1.45, 1.95, 2.17, 2.76 **GEN** = 1.26; 0.95, 1.77, 2.39, 2.92
		PFC41 (item 88)—"Are you able to sit down in and stand up from a low, soft couch?"	$R^2_{12}=.026$ $R^2_{23}=.000$	OA = 2.94; − 2.07, − 1.14, − 0.62, 0.13 **GEN** = 2.50; − 2.48, − 1.64, − 1.12, − 0.28
		PFC45r1 (item 89)[a]—"Are you able to sit on and get up from the toilet?"	$R^2_{12}=.023$ $R^2_{23}=.000$	OA = 3.35; − 2.94, − 1.92, − 1.45, − 0.77 **GEN** = 3.12; − 2.69, − 2.22, − 1.81, − 1.19
		PFC47 (item 91)—"Are you able to be out of bed most of the day?"	$R^2_{12}=.027$ $R^2_{23}=.007$	OA = 1.88; − 2.79, − 2.43, − 2.22, − 1.80 **GEN** = 1.18; − 3.51, − 2.78, − 2.07, − 1.38
		PFC49 (item 92)—"Are you able to water a house plant?"	$R^2_{12}=.022$ $R^2_{23}=.009$	OA = 3.79; − 2.77, − 2.50, − 2.24, − 1.83 **GEN** = 2.45; − 3.10, − 2.60, − 2.14, − 1.63
		PFB50 (item 99)—"How much difficulty do you have doing your daily physical activities, because of your health?"	$R^2_{12}=.023$ $R^2_{23}=.004$	**OA** = 4.24; − 1.67, − 0.90, − 0.08 GEN = 3.03; − 1.53, − 0.66, 0.28
		PFC56 (item 103)—"Does your health now limit you in walking about the house?"	$R^2_{12}=.031$ $R^2_{23}=.000$	**OA** = 4.29; − 2.19, − 1.56 GEN = 3.66; − 1.85, − 1.27
		PFB48 (item 111)—"Does your health now limit you in taking a shower?"	$R^2_{12}=.023$ $R^2_{23}=.008$	**OA** = 4.79; − 2.12, − 1.59 GEN = 3.07; − 2.03, − 1.39
		PFB49 (item 112)—"Does your health now limit you in going for a short walk (less than 15 min)?"	$R^2_{12}=.047$ $R^2_{23}=.001$	**OA** = 3.24; − 2.37, − 1.81, − 1.34 GEN = 2.95; − 1.94, − 1.39, − 0.81

Table 2 (continued)

PROMIS item banks	Populations compared	Items with DIF	McFadden R^2 values	Slope; and thresholds parameters
		PFB54 (item 115)—“Does your health now limit you in going OUTSIDE the home, for example to shop or visit a doctor’s office?”	R^2_{12} = .020 R^2_{23} = .015	OA = 3.99; − 2.70, − 1.89, − 1.38 **GEN** = 2.23; − 3.25, − 1.90, − 1.20
	Physiotherapy (PT) vs. general population (GEN)	PFA13 (item 6)—“Are you able to exercise for an hour?”	R^2_{12} = .023 R^2_{23} = .001	**PT** = 3.00; − 1.82, − 1.22, − 0.81, − 0.11 GEN = 2.36; − 1.49, − 0.96, − 0.43, 0.35
		PFA51 (item 38)[a]—“Are you able to sit on the edge of a bed?”	R^2_{12} = .108 R^2_{23} = .000	PT = 3.10; − 1.71, − 1.38, − 0.95 **GEN** = 2.86; − 2.81, − 2.25, − 1.89
		PFC32 (item 83)—“Are you able to climb up 5 flights of stairs?”	R^2_{12} = .021 R^2_{23} = .000	**PT** = 2.05; − 2.12, − 1.58, − 1.24, − 0.64 GEN = 2.16; − 1.82, − 1,25, − 0.75, − 0.04
		PFB49 (item 112)—“Does your health now limit you in going for a short walk (less than 15 min)?”	R^2_{12} = .032 R^2_{23} = .003	**PT** = 4.01; − 2.22, − 1.77, − 1.46 GEN = 3.21; − 2.07, − 1.56, − 1.03
Pain interference	Chronic pain (PAIN) vs. general population (GEN)	PAININ20 (item 10)—“In the past 7 days... How much did pain feel like a burden to you...”	R^2_{12} = .023 R^2_{23} = .000	**PAIN** = 2.92; − 1.56, − 0.59, 0.07, 1.19 GEN = 2.83; − 1.18, − 0.12, 0.53, 1.64
	Osteoarthritis (AMS-OA) vs. general population (GEN)	PAININ42 (item 30)—“In the past 7 days... How often did pain prevent you from standing for more than one hour...”	R^2_{12} = .026 R^2_{23} = .000	**AMS-OA** = 2.29; − 0.97, − 0.40, 0.30, 0.95 GEN = 2.23; − 0.37, 0.18, 0.80, 1.45
Pain behavior	Chronic Pain (PAIN) vs. general population (GEN)	PAINBE45 (item 28)[b]—“In the past 7 days... When I was in pain I thrashed...”	R^2_{12} = .021 R^2_{23} = .000	PAIN = 2.11; 1.21, 1.79, 2.51 **GEN** = 2.17; 0.59, 1.33, 2.20
	Rheumatoid Arthritis (RA) vs. general population (GEN)	PAINBE24 (item 12)[b]—“In the past 7 days... When I was in pain I moved stiffly...”	R^2_{12} = .029 R^2_{23} = .000	**RA** = 1.86; − 1.48, − 0.58, 0.73, 2.22 GEN = 1.60; − 0.60, 0.16, 1.31, 2.99
		PAINBE25 (item 13)[b]—“In the past 7 days... When I was in pain, I called out for someone to help me...”	R^2_{12} = .026 R^2_{23} = .001	**RA** = 1.53; − 0.41, 0.62, 2.13, 3.54 GEN = 1.81; 0.36, 1.23, 2.36, 3.68

The bold marked populations in the ‘slopes and thresholds column’ means that the population had lower thresholds compared to the other population. Indicating that at the same level of the domain (same level of physical function, pain interference, or pain behavior), the population with lower thresholds will endorse higher item response categories

[a] Present in the PROMIS-PF 20a item short form

[b] Present in the PROMIS-PB 7a item short form

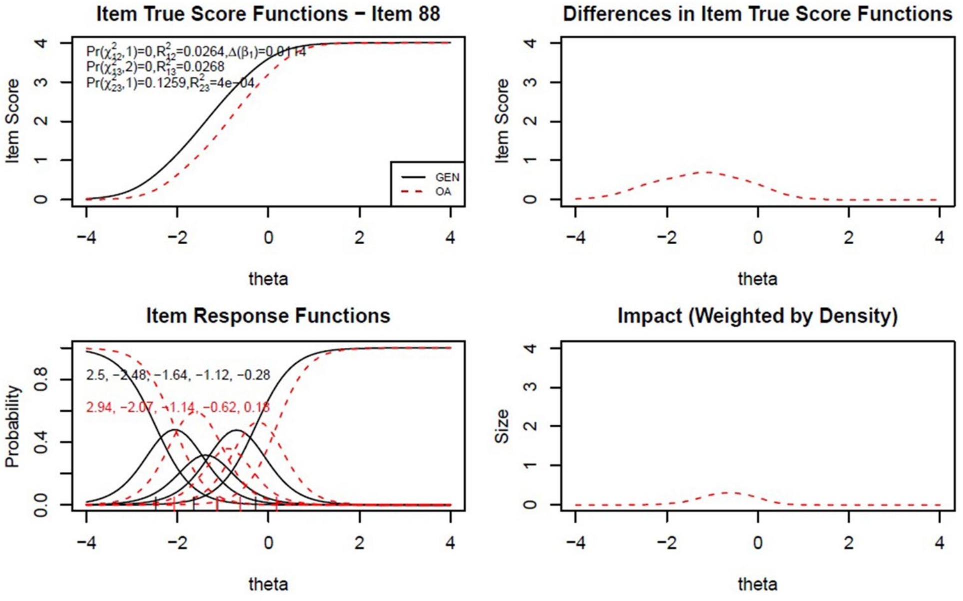


Fig. 2 The item characteristics plot of item PFC41 (item 88)—“Are you able to sit down in and stand up from a low, soft couch?” from the PROMIS-PF bank for the comparison OA vs. GEN, includes four plots: (1) the ICCs or item true score functions per population (OA vs. GEN), illustrates which population has higher item scores given the levels of theta; (2) the absolute difference between the ICCs or differences in item true score functions, showing the difference in item scores given the levels of theta; (3) the item response functions, including the item slope and threshold parameters, per population, visualizing which population has higher probabilities of endorsing the response categories at issue given the levels of theta; and (4) the impact weighted by density, showing the absolute difference in item scores weighted by the theta distribution of the samples. In this example, all four plots show negligible impact of DIF

DIF. Appendices 1–3 show the item characteristics plots and TCCs for the DIF items found in this study per bank.

The interpretation of the DIF is illustrated for the comparison of the OA population with the GEN population for item PFC41 (item 88) of the PROMIS-PF bank: “Are you able to sit down in and stand up from a low, soft couch?” The McFadden’s pseudo R^2-change value for the difference between models 2 and 3 was below the criterion of 0.02 ($R^2{}_{23}=.0004$), indicating no non-uniform DIF, but the McFadden’s pseudo R^2-change value for the difference between models 1 and 2 was above the criterion of 0.02 ($R^2{}_{12}=.0264$), indicating uniform DIF (Fig. 2, left upper plot). The threshold parameters for the OA population (−2.07, −1.14, −0.62, 0.13) were slightly higher than for the GEN population (−2.48, −1.64, −1.12, −0.28), indicating that the OA population will endorse lower response categories at the same level of physical function (Fig. 2, left lower plot). For the interpretation of item PFC41 this means that at the same level of physical function, the OA population is less likely to be able to sit down in and stand up from a low, soft couch than the GEN population.

The overall impact of DIF on the item scores and *T*-scores for all item banks was negligible. For example, the item characteristics plots for item PFC41 showed a small difference between the item true score functions (Fig. 2, left upper plot), indicating that the difference in item score given the levels of theta was minimal. In addition, the TCC of the PROMIS-PF bank, comparing the OA and GEN populations, showed that the area between the curves was negligible in both the left and right plot, indicating minimal impact of DIF by population on test scores (see Fig. 3). Similar results were found for the other comparisons and banks of this study.

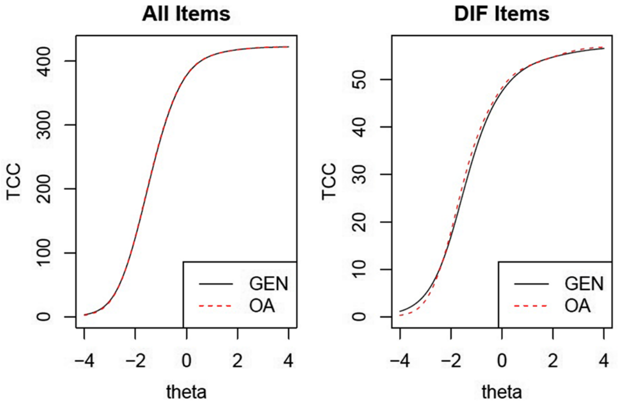


Fig. 3 The test characteristic curves (TCCs) of the PROMIS-PF bank of the comparison OA vs. GEN, shows the test scores (raw summary score) for all 121 PROMIS-PF items (ignoring DIF) per population in the left plot, and the scores per population for only the 14 items having DIF in the right plot. The area between the two curves within one plot provides an indication of the impact of DIF on the test score, showing in this example negligible impact of DIF

Only three items of the PROMIS-PF bank, items PFA51 (OA vs. PT), PFC33r1, and PFC40 (OA vs. AMS-PAIN), showed relatively high (> 0.05) R^2-change values (R^2_{12} = 0.220; 0.065; 0.052, respectively). The item characteristic plots of PFA51 and PFC40 showed relatively large differences in item true score functions and item response functions, and the item characteristic plots of PFC33r1 showed clustered item response functions (see Online Appendix 1). However, the corresponding TCCs indicate that the impact of DIF of these items was minimal on the test scores (see Online Appendix 1).

Discussion

Our aim was to investigate the validity of comparisons across populations of patients with different musculoskeletal disorders and persons from the general population when applying the PROMIS-PF, PROMIS-PI, and PROMIS-PB banks, by evaluating DIF across (patient) populations. We found some items with DIF, but the magnitude and impact of DIF on the *T*-scores were negligible, supporting the universal applicability of the item banks. The item banks can be used by health providers and clinical researchers to compare patients with different musculoskeletal disorders and healthy persons.

Although the impact of the DIF items on the *T*-scores was negligible in the current study, there was a possible explanation for some DIF items. For the PROMIS-PF bank, the two comparisons that showed the largest amount of DIF items were the comparisons between the patients with hip or knee OA and chronic pain, and between the patients with hip or knee OA and the persons from the general population. In both comparisons, 14 items with DIF were found. For 12 out of 14 items with DIF, patients with hip or knee OA were less likely to endorse the items than patients with chronic pain, given the same level of theta. For instance, at the same level of function, patients with hip or knee OA were slightly less likely to be able to run 100 yards, to run two, five, or ten miles, to get up from or kneel on the floor, to squat and get up, to take a tub bath, and to sit on and get up from a low couch or toilet. At the same level of function, patients with hip and knee OA were less likely to endorse that they were able to squat and get up, to get in and out of a car, to run ten miles, to sit down in and stand up from a couch or toilet, and to be out of bed most of the day, than persons from the general population. Moreover, patients with hip or knee OA were more likely to endorse to have more difficulty in doing daily physical activities and to have more limitations in walking around the house, taking a shower, going for a short walk, and going outside the home, than persons from the general population given the same level of theta. All these DIF results may be explained by the fact that the

Table 3 Number of items with DIF in previous studies of the PROMIS-PF, PROMIS-PI, and PROMIS-PB banks

Study	Study population	DIF for					
		Age	Gender	Education	Administration mode	Disease activity	Language
PROMIS-PF							
Hung et al. [22]	Orthopedic patients	56	16	10	–	–	–
Rose et al. [10]	General population and clinical samples	4	7	1	–	–	–
Hung et al. [36]	Spi-l disorders	45	28	13	–	–	–
Oude Voshaar et al. [15]	Rheumatoid arthritis	5	7	–	–	–	25 (Dutch vs. English)
Paz et al. [23, 37]	General population	30	–	–	–	–	50 (Spanish vs. English)
Crins et al. [16]	Chronic pain	1	0	–	0	–	4 (Dutch vs. English)
Crins et al. [17]	Physiotherapy patients	2	14	–	–	–	–
PROMIS-PI							
Amtmann et al. [11]	General population and clinical samples	8	1	0	–	–	–
Crins et al. [14]	Chronic pain	0	0	–	0	–	2 (Dutch vs. English)
Paz et al. [18]	General population	–	–	–	–	–	1 (Spanish vs. English)
Crins et al. [19]	Rheumatoid arthritis	0	0	–	0	0	0 (Dutch vs. Flemish)
PROMIS-PB							
Revicki et al. [12]	General population and clinical samples	5	1	0	–	–	–
Crins et al. [13]	Chronic pain	0	0	–	0	–	6 (Dutch vs. English)
Crins et al. [19]	Rheumatoid arthritis	0	1	–	0	0	1 (Dutch vs. Flemish)

'–' not applicable

activities addressed in these items are specifically influenced by knee and hip problems.

With respect to the PROMIS-PB bank, patients with RA were more likely to endorse that they moved stiffly when were in pain, than persons from the general population given the same level of theta. This may be because stiffness is one of the typical clinical characteristics of RA.

For the Dutch-Flemish PROMIS-PF, PROMIS-PI, and PROMIS-PB banks, the validity of comparisons across populations has also been shown for other comparisons populations. Previous studies showed no DIF or DIF with negligible impact on the *T*-score, for sub-populations differing in age, gender, education level, administration mode (paper–pencil vs. web-based), disease activity or language. Results on DIF in those studies are summarized in Table 3. The results of the current study, addressing the Dutch-Flemish PROMIS banks, can most likely be generalized to the original American-English PROMIS banks, as previous studies of the Dutch-Flemish PROMIS-PF, PROMIS-PI, and PROMIS-PB banks showed the absence of DIF or DIF with negligible impact between the Dutch and English language [13, 14, 16]. The current results combined with the previous results of studies on DIF for other variables indicate that the item parameters seem to be quite stable across different (sub)populations. Only the PROMIS-PF bank maybe needs more caution. For instance, the DIF found for items PFA51, PFC33r1, and PFC40 of the PROMIS-PF bank might be of a slight concern.

Although it does not seem that there is a high impact of items with DIF on the item bank *T*-scores, the impact on short-form *T*-scores or CAT *T*-scores might be larger since only a small number of items are administered. It could be that from these small numbers of items in short forms and CATs, just those items with DIF are included, and that the DIF is cumulative. From the items with DIF found in the current study, three items (PFA51, PFA56, and PFC45r1) from the PROMIS-PF bank and 3 items (PAINBE24, PAINBE25, PAINBE45) from the PROMIS-PB bank are present in, respectively, the PROMIS-PF 20a-item short form and the PROMIS-PB 7a-item short form. The impact of DIF on short-form *T*-scores and CAT *T*-scores could be examined in a future study.

Study strengths are that we were able to use large and diverse datasets. However, for future research it might be important to include patients with disorders that differ considerable from patients with musculoskeletal disorders (e.g., patients with heart condition, cancer or stroke). A study limitation is that we used the logistic regression method to detect DIF and the McFadden's pseudo R^2 change of 2% as critical value only, while there are multiple methods for detection of DIF and multiple criteria available [33]. We chose our method and critical value because these are commonly applied in PROMIS studies [7, 13, 15, 18, 23, 30, 33]. For future studies, we recommend to study and compare other methods and cut-off values as well, for instance the Monte Carlo simulations approach which facilitates empirical identification of the critical R^2-change value [31]. Future studies could also consider to use, as an alternative approach, multiple group DIF analysis which enables to compare multiple clinical groups and a reference groups simultaneously. A disadvantage, however, of this alternative approach, is that it provides an overall test for DIF between any of the groups only and, therefore, offers less insight in the differences between each group.

In conclusion, this study contributes to the evidence for the universal applicability of PROMIS across (patient) populations. Moreover, our results provide evidence that comparisons across patients with different musculoskeletal disorders and persons from the general population are valid, when applying the PROMIS-PF, PROMIS-PI, and PROMIS-PB banks.

Acknowledgements The Dutch-Flemish PROMIS group is an initiative that aims to translate, investigate, and implement PROMIS item banks, short forms, and CATS in the Netherlands and Flanders (http://www.dutchflemishpromis.nl). The CHECK-cohort study is funded by the Dutch Arthritis Foundation. Involved are Erasmus Medical Center Rotterdam; Kennemer Gasthuis Haarlem; Leiden University Medical Center; Maastricht University Medical Center; Martini Hospital Groningen/Allied Health Care Center for Rheumatology and Rehabilitation Groningen; Medical Spectrum Twente Enschede/Ziekenhuisgroep Twente Almelo; Reade Center for Rehabilitation and Rheumatology; St. Maartenskliniek Nijmegen; University Medical Center Utrecht; and Wilhelmina Hospital Assen.

Compliance with ethical standards

Conflict of interest Dr. C.B. Terwee is president of the PROMIS Health Organization. All authors are members of the Dutch-Flemish PROMIS group. All authors have no financial or non-financial conflicts of interest.

Ethical approval As this study did not involve experiments with patients, it was exempt from ethical approval according to the Dutch Medical Research in Human Subjects Act (WMO) by the medical ethics committee of the VU University Medical Center, and of Slotervaart and Reade, Amsterdam.

Human and animal rights All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and national research committees and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent According to the Dutch Medical Research in Human Subjects Act (WMO), obtaining informed consent was not necessary. The waiver includes that a signature from the patient on a separate informed consent form was not required, but that completion of the questionnaire by the patient can be considered as informed consent.

References

1. Cella, D., Gershon, R., Lai, J., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*(1), 133–141.
2. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., … 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. *Journal of Clinical Epidemiology, 63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011.
3. Fries, J. F., Bruce, B., Bjorner, J., & Rose, M. (2006). More relevant, precise, and efficient items for assessment of physical function and disability: Moving beyond the classic instruments. *Annals of the Rheumatic Diseases, 65*(Suppl 3), iii16–i21. https://doi.org/10.1136/ard.2006.059279.
4. Fries, J., Rose, M., & Krishnan, E. (2011). The PROMIS of better outcome assessment: Responsiveness, floor and ceiling effects, and internet administration. *The Journal of Rheumatology, 38*(8), 1759–1764. https://doi.org/10.3899/jrheum.110402.
5. Fries, J. F., Krishnan, E., Rose, M., Lingala, B., & Bruce, B. (2011). Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Research & Therapy, 13*(5), R147. https://doi.org/10.1186/ar3461.
6. Oude Voshaar, M. A. H., ten Klooster, P. M., Glas, C. A. W., Vonkeman, H. E., Krishnan, E., & van de Laar, M. A. F. J. (2014). Relative Performance of Commonly Used Physical Function Questionnaires in rheumatoid arthritis and a patient-reported outcomes measurement information system computerized adaptive test. *Arthritis & Rheumatology, 66*(10), 2900–2908. https://doi.org/10.1002/art.38759.
7. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., … Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5), 22–31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.
8. Holland, P., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
9. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
10. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526. https://doi.org/10.1016/j.jclinepi.2013.10.024.
11. Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W.-H., Choi, S., Revicki, D., … Lai, J.-S. (2010). Development of a PROMIS item bank to measure pain interference. *Pain, 150*(1), 173–182. https://doi.org/10.1016/j.pain.2010.04.025.
12. Revicki, D. A., Chen, W.-H., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., … Keefe, F. J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain, 146*(1–2), 158–169. https://doi.org/10.1016/j.pain.2009.07.029.
13. Crins, M. H. P., Roorda, L. D., Smits, N., de Vet, H. C. W., Westhovens, R., Cella, D., … Terwee, C. B. (2016). Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank in patients with chronic pain. *European Journal of Pain, 20*(2), 284–296. https://doi.org/10.1002/ejp.727.
14. Crins, M. H. P., Roorda, L. D., Smits, N., de Vet, H. C. W., Westhovens, R., Cella, D., … Terwee, C. B. (2015). Calibration and validation of the Dutch-Flemish PROMIS pain interference item bank in patients with chronic pain. *PLoS ONE, 10*(7), 1–18. https://doi.org/10.1371/journal.pone.0134094.
15. Oude Voshaar, M. A. H., ten Klooster, P. M., Glas, C. A. W., Vonkeman, H. E., Taal, E., Krishnan, E., … van de Laar, M. A. F. J. (2014). Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. *PLoS ONE, 9*(3), e92367. https://doi.org/10.1371/journal.pone.0092367.
16. Crins, M. H. P., Terwee, C. B., Klausch, T., Smits, N., de Vet, H. C. W., Westhovens, R., … Roorda, L. D. (2017). The Dutch-Flemish PROMIS physical function item bank exhibited strong psychometric properties in patients with chronic pain. *Journal of Clinical Epidemiology, 87*, 47–58. https://doi.org/10.1016/j.jclinepi.2017.03.011.
17. Crins, M., Wees, van der, P., Klausch, T., Dulmen, van, S., Roorda, L., & Terwee, C. (2018). Psychometric properties of the PROMIS physical function item bank in patients receiving physical therapy. *PLoS ONE, 13*(2), 1–14. https://doi.org/10.1371/journal.pone.0192187.
18. Paz, S. H., Spritzer, K. L., Reise, S. P., & Hays, R. D. (2017). Differential item functioning of the patient-reported outcomes information system (PROMIS®) pain interference item bank by language (Spanish versus English). *Quality of Life Research, 26*(6), 1451–1462. https://doi.org/10.1007/s11136-017-1499-3.
19. Crins, M., Terwee, C., Westhovens, R., Schaardenburg van, D., Smits, N., Joly, J., … Roorda, L. (2018). First validation of the full PROMIS pain interference and pain behavior item banks in patients with rheumatoid arthritis. *Arthritis Care & Research* (**submitted for publication**).
20. Schuller, W., Terwee, C., Klausch, T., Roorda, L., Rohrich, D., Ostelo, R., … de Vet, H. (2018). Validation of the Dutch-Flemish PROMIS pain behavior item bank in patients with musculoskeletal complaints. *PLoS ONE* (**submitted for publication**).
21. Schuller, W., Terwee, C., Klausch, T., Roorda, L., Rohrich, D., Ostelo, R., … de Vet, H. (2018). Validation of the Dutch-Flemish PROMIS pain interference item bank in patients with musculoskeletal complaints. *Spine*. https://doi.org/10.1097/BRS.0000000000002847.
22. Hung, M., Clegg, D. O., Greene, T., & Saltzman, C. L. (2011). Evaluation of the PROMIS physical function item bank in orthopaedic patients. *Journal of Orthopaedic Research, 29*(6), 947–953. https://doi.org/10.1002/jor.21308.
23. Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2013). Evaluation of the Patient-Reported Outcomes Information System (PROMIS(®)) Spanish-language physical functioning items. *Quality of Life Research, 22*(7), 1819–1830. https://doi.org/10.1007/s11136-012-0292-6.
24. Wesseling, J., Boers, M., Viergever, M. A., Hilberdink, W. K. H. A., Lafeber, F. P. J. G., Dekker, J., & Bijlsma, J. W. J. (2016). Cohort Profile: Cohort Hip and Cohort Knee (CHECK) study. *International Journal of Epidemiology, 45*(1), 36–44. https://doi.org/10.1093/ije/dyu177.
25. Wesseling, J., Dekker, J., Van Den Berg, W. B., Bierma-Zeinstra, S. M. A., Boers, M., Cats, H. A., … Bijlsma, J. W. J. (2009). CHECK (Cohort Hip and Cohort Knee): Similarities and differences with the osteoarthritis initiative. *Annals of the Rheumatic Diseases, 68*(9), 1413–1419. https://doi.org/10.1136/ard.2008.096164.
26. Terwee, C. B., Roorda, L. D., de Vet, H. C. W., Dekker, J., Westhovens, R., van Leeuwen, J., … Boers, M. (2014). Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). *Quality of Life Research, 23*(6), 1733–1741. https://doi.org/10.1007/s11136-013-0611-6.
27. PROMIS. (2017). *PROMIS pain interference scoring manual*. Retrieved November 29, 2017, from http://www.healthmeasures.

net/images/PROMIS/manuals/PROMIS_Pain_Interference_Scoring_Manual_02232017.pdf.

28. PROMIS. (2015). *PROMIS pain behavior scoring manual*. Retrieved November 29, 2017, from http://www.healthmeasures.net/images/promis/manuals/PROMIS_Pain_Behavior_Scoring_Manual.pdf.
29. Cook, K. F., Keefe, F., Jensen, M. P., Roddey, T. S., Callahan, L. F., Revicki, D., … Amtmann, D. (2013). Development and validation of a new self-report measure of pain behaviors. *Pain, 154*(12), 2867–2876. https://doi.org/10.1016/j.pain.2013.08.024.
30. Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S115–S123. https://doi.org/10.1097/01.mlr.0000245183.28384.ed.
31. 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. *Journal of Statistical Software, 39*(8), 1–30.
32. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016). *Logistic ordinal regression differential item functioning using IRT, version 0.3-3*. Retrieved December 14, 2016, from https://cran.r-project.org/web/packages/lordif/lordif.pdf.
33. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. A. G. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes, 8*, 1–9. https://doi.org/10.1186/1477-7525-8-81.
34. PROMIS. (2014). *Minimum requirements for the release of PROMIS instruments after translation and recommendations for further psychometric evaluation*. Retrieved June 2, 2018, from http://www.healthmeasures.net/images/PROMIS/Standards_for_release_of_PROMIS_instruments_after_translation_v8.pdf.
35. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates. https://doi.org/10.1016/C2013-0-10517-X.
36. Hung, M., Hon, S. D., Franklin, J. D., Kendall, R. W., Lawrence, B. D., Neese, A., … Brodke, D. S. (2014). Psychometric properties of the promis physical function item bank in patients with spinal disorders. *Spine, 39*(2), 158–163. https://doi.org/10.1097/BRS.0000000000000097.
37. Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2013). Age-related differential item functioning for the Patient-Reported Outcomes Information System (PROMIS®) physical functioning items. *Primary Health Care: Open Access*. https://doi.org/10.4172/2167-1079.1000131.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.