First Validation of the Full PROMIS Pain Interference and Pain Behavior Item Banks in Patients With Rheumatoid Arthritis

Crins, M.H.P.; Terwee, C.B.; Westhovens, R.; van Schaardenburg, D.; Smits, N.; Joly, J.; Verschueren, P.; Van der Elst, K.; Dekker, J.; Boers, M.; Roorda, L.D.

DOI
10.1002/acr.24077

Publication date
2020

Document Version
Final published version

Published in
Arthritis Care & Research

American College of Rheumatology
Empowering Rheumatology Professionals

Arthritis Care & Research
Vol. 72, No. 11, November 2020, pp 1550–1559
DOI 10.1002/acr.24077

First Validation of the Full PROMIS Pain Interference and Pain Behavior Item Banks in Patients With Rheumatoid Arthritis

Martine H. P. Crins,[1] Caroline B. Terwee,[2] Rene Westhovens,[3] Dirkjan van Schaardenburg,[4] Niels Smits,[5] Johan Joly,[3] Patrick Verschueren,[3] Kristien Van der Elst,[3] Joost Dekker,[6] Maarten Boers,[7] and Leo D. Roorda[1]

Objective. Pain interference and pain behavior are highly relevant outcomes in patients with rheumatoid arthritis (RA). The Patient-Reported Outcomes Measurement Information System (PROMIS) is a universally applicable set of item banks measuring patient-reported health, and if applied as computerized adaptive tests (CATs), more efficiently and precisely than current instruments. The objective was to study the psychometric properties of the Dutch-Flemish PROMIS pain interference (PROMIS-PI) and the PROMIS pain behavior (PROMIS-PB) item banks in patients with RA.

Methods. A total of 2,029 patients with RA completed the full PROMIS-PI (version 1.1, 40 items), and 1,554 patients completed the full PROMIS-PB (version 1.1, 39 items). The following psychometric properties were studied: unidimensionality, local dependence, monotonicity and graded response model (GRM) fit, cross-cultural validity (differential item functioning [DIF] for language [Dutch versus Flemish]), other forms of measurement invariance, construct validity, reliability, and floor and ceiling effects.

Results. The PROMIS-PI and PROMIS-PB banks were sufficiently unidimensional (Omega-hierarchical [Omega-H] 0.99, 0.95, and explained common variance 0.95, 0.78, respectively), had negligible local dependence (0.3–1.4% of item pairs), good monotonicity (H 0.75, 0.46), and a good GRM model fit (no misfitting items). Furthermore, both item banks showed good cross-cultural validity (no DIF for language), measurement invariance (no DIF for age, sex, administration mode, and disease activity), good construct validity (all hypotheses met), high reliability (>0.90 in the range of patients with RA), and an absence of floor and ceiling effects (0% minimum or maximum score, respectively).

Conclusion. Both PROMIS-PI and PROMIS-PB banks showed good psychometric properties in patients with RA and can be used as CATs in research and clinical practice.

INTRODUCTION

Rheumatoid arthritis (RA) is characterized by pain and swelling of joints leading to disability and has a considerable impact on quality of life if not sufficiently treated (1–3). The patient perspective is key to assess the outcomes of treatment of RA. The Outcome Measures in Rheumatology initiative developed a core set of outcomes for RA, including level of pain, physical function, and fatigue (4–6). The International Consortium for Health Outcomes Measurement (ICHOM) added emotional health to this core set (7). ICHOM recommends pain interference as a relevant outcome domain (7). A recent study, on standardizing and personalizing the treat-to-target approach for RA, highlighted the importance of pain interference because patients with RA selected the domain pain interference, in addition to physical function, as their highest priority outcome domain (8). Pain interference is defined as the degree to which pain interferes with or limits an individual's social, mental, and physical activities. A related and additional important aspect of pain is pain behavior, defined as behaviors that typically indicate to others that one is experiencing pain (9–11). Both constructs

Supported by the Virtutis Opus Foundation.
[1]Martine H. P. Crins, MSc, Leo D. Roorda, MD, PT, PhD: Amsterdam Rehabilitation Research Center, Reade, Amsterdam, The Netherlands; [2]Caroline B. Terwee, PhD: Amsterdam UMC, Vrije Universiteit Amsterdam, and Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; [3]Rene Westhovens, PhD, Johan Joly, MSc, Patrick Verschueren, PhD, Kristien Van der Elst, MSc: KU Leuven and University Hospitals Leuven, Leuven, Belgium; [4]Dirkjan van Schaardenburg, PhD: Amsterdam Rheumatology and Immunology Center, Reade, and Academic Medical Center, Amsterdam, The Netherlands; [5]Niels Smits, PhD: University of Amsterdam, Amsterdam, The Netherlands; [6]Joost Dekker, PhD: Amsterdam University Medical Centers, Amsterdam, The Netherlands; [7]Maarten Boers, PhD: Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, Amsterdam Rheumatology and Immunology Center, Reade, and VU University Medical Center, Amsterdam, The Netherlands.

No potential conflicts of interest relevant to this article were reported.

Address correspondence to Martine H. P. Crins, MSc, Reade, Dr. Jan van Breemenstraat, Postbus 58271, 1040 HG Amsterdam, The Netherlands. Email: martinecrins@hotmail.com.

Submitted for publication May 15, 2019; accepted in revised form September 24, 2019.

SIGNIFICANCE & INNOVATIONS

- This is the first extensive study worldwide of the psychometric properties of the PROMIS pain interference and PROMIS pain behavior banks in patients with rheumatoid arthritis (RA), and it provides evidence that these banks can be applied in these patients.
- Short forms, derived from these banks, yield valid and standardized results and allow routine monitoring of pain interference and behavior in groups of patients with RA, both in clinical practice and research.
- Computerized adaptive tests, derived from these banks, yield valid, precise, and standardized results, are efficient, user-friendly, and feasible, with little administration time, and they allow routine monitoring of pain interference and behavior in individuals and groups of patients with RA, both in clinical practice and research.

(domains) can be considered important to measure in patients with RA.

Problems with current Patient-Reported Outcome Measures (PROMs) include practical burden for patients, irrelevance of some items, measurement quality (i.e., measurement imprecision), and limitations in measurement range (i.e., ceiling and floor effects) (12). Results can be difficult to interpret, cannot be compared between the many different existing PROMs, and often cannot be used to compare patients with different diseases (13).

The Patient-Reported Outcomes Measurement Information System (PROMIS) is an innovative set of instruments to measure patient-reported health with the potential to overcome the shortcomings of existing PROMs (14). PROMIS instruments are developed based on item response theory (IRT) methods and consist of so-called item banks. An item bank is a set of items (questions) that measure a single construct, such as pain interference (15). Because the items of a bank are calibrated to an IRT model, the items can be administered as short forms, fixed length subsets of items out of the bank, or highly efficient computerized adaptive tests (CATs), a computer-administered measure in which successive items are selected by a computer algorithm informed by responses to previous items (16,17). In a CAT, individuals only have to complete a small number of highly informative and relevant items (15,17,18). PROMIS is intended to be universally applicable and therefore usable in and across different patient populations.

Short forms and CATs derived from the PROMIS pain interference (PROMIS-PI) and PROMIS pain behavior (PROMIS-PB) banks have already been applied in patients with RA (19,20). These applications in patients with RA are based on the assumption that PROMIS banks are indeed universally applicable. However, sufficient psychometric properties of these banks have not yet been demonstrated for this patient group. Therefore, the aim of this study was to examine the psychometric properties of the PROMIS-PI and PROMIS-PB in patients with RA.

PATIENTS AND METHODS

Study participants. For this study, 2,489 Dutch patients (from Reade, Center for Rehabilitation and Rheumatology, Amsterdam, The Netherlands) and 850 Flemish patients (from University Hospitals Leuven, Leuven, Belgium) were invited. Patients were eligible if they were age ≥18 years, had RA according to the American College of Rheumatology criteria, and provided informed consent (21).

Procedures. The study was approved by both the Dutch and Belgian local institutional review boards. The Dutch patients were invited by email or letter, and the Flemish patients by their treatment team, to complete a web-based (digital) or paper questionnaire.

Measures. The questionnaire included questions addressing demographic and clinical characteristics. The Disease Activity Score in 28 joints (DAS28) was derived from the medical record. The questionnaire also included the full versions of the Dutch-Flemish PROMIS-PI (version 1.1) and PROMIS-PB (version 1.1) banks (22). Flemish strongly resembles Dutch and is spoken in Flanders, a part of Belgium. The PROMIS-PI bank assesses self-reported consequences of pain on relevant aspects of life. This includes the extent to which pain hinders engagement with social, cognitive, emotional, physical, and recreational activities (23). The bank consists of 40 items. The time frame is the past 7 days, and the bank uses 3 different 5-point Likert response scales, in which high scores represent more of the construct (10,11,23,24). The PROMIS-PB bank measures self-reported external manifestations of pain, behaviors that typically indicate to others that an individual is experiencing pain (24). The bank contains 39 items. Patients rate how frequently they have engaged in the different types of pain behavior in the past 7 days on a 6-point Likert response scale, in which high scores represent more of the construct (11,25). Comparable to previous studies of this bank, this study excluded patients who rated the "had no pain" response category on any of the items, resulting in 5 response options for the IRT analyses (25,26). Good psychometric properties of the PROMIS-PI and PROMIS-PB banks have already been found in patients with chronic pain (25,27).

The study questionnaire also contained 4 legacy instruments, 2 generic and 2 RA disease-specific. The 2 generic instruments were the pain intensity item (Global07), from the Dutch-Flemish PROMIS version 1.2 global health item bank, and the 10 items of the Short Form 36 health survey physical functioning scale (SF-36 PF10) (28–30). The 2 RA disease-specific PROMs were the Health Assessment Questionnaire disability index (HAQ DI) and the Rheumatoid Arthritis Impact of Disease (RAID) questionnaire (31–34). For the PROMIS global health pain intensity item (range

Table 1. Psychometric properties, research questions, analysis, statistics, criteria, software packages, references, and results of the PROMIS-PI and PROMIS-PB item banks in patients with rheumatoid arthritis*

Psychometric properties, research questions, and analysis, with statistic	Criterion	Software package	Refs.	PROMIS-PI results (40 items)		PROMIS-PB results (39 items)	
				Unscaled	Scaled	Unscaled	Scaled
IRT model assumptions, fit, and descriptives†							
IRT model assumptions							
1. Unidimensionality: do the items assess 1 and only 1 construct?	–	R, Lavaan 0.5-16	44	–	–	–	–
Confirmatory factor analysis							
Comparative fit index	>0.95	–	45	0.997	0.973	0.980	0.855
Tucker-Lewis Index	>0.95	–	45	0.996	0.972	0.978	0.847
Root mean square error of approximation	<0.06	–	45	0.128	0.113	0.085	0.093
Explanatory factor analysis							
Variance explained by 1st factor	>20%	–	18	79%	–	50%	–
Ratio variance explained by 1st/2nd factor	>4	–	18	29.5	–	13.2	–
Confirmatory bifactor analysis							
Omega-hierarchical	>0.80	–	46	0.99	–	0.95	–
Explained common variance	>0.60	–	46	0.95	–	0.78	–
Comparative fit index	>0.95	–	45	0.998	0.981	0.993	0.941
Tucker-Lewis Index	>0.95	–	45	0.998	0.980	0.992	0.935
Root mean square error of approximation	<0.06	–	45	0.092	0.095	0.051	0.060
2. Local independence: are the items only related to the construct being measured and not to other factors?	–	R, Lavaan 0.5-16	44	–	–	–	–
Residual correlation matrix							
Percentage of item pairs with correlations >0.20	–	–	18	0.3%	–	1.4%	–
Impact on IRT-parameters of removing most locally dependent items							
Mean and maximum difference in item slopes	–	–	–	0.015, 0.054	–	0.000, 0.057	–
Mean and maximum difference in item thresholds	–	–	–	−0.004, −0.048	–	−0.001, −0.108	–
3. Monotonicity: does the probability of an affirmative response to the items increase with increasing levels of the underlying construct?	–	R, Mokken	47	–	–	–	–
Mokken scale analysis							
Scalability coefficients of all item pairs	>0	–	18,47	all items >0	–	all items >0	–
Scalability coefficients of the items in relation to the scale	≥0.3	–	18,47	all items ≥0.3	–	all items ≥0.3	–
Scalability coefficient of the scale	≥0.3	–	18,47	0.75	–	0.46	–
IRT-model fit	–	R, Mirt 3.3.2	48	–	–	–	–
Graded response model fit: can the relationships between the items adequately be described with this IRT model?							
S-X2 statistic, *P* value per item	>0.001	–	18	100%	–	100%	–
IRT-model descriptives							
Item parameters: what are the mean and range of the item parameters?							
Item slope							
Mean	–	–	–	4.2	–	1.9	–
Range	–	–	–	2.1–7.0	–	1.4–2.8	–
Item thresholds							
Range	–	–	–	−0.4, 3.4	–	−1.6, 5.4	–

(Continued)

Table 1. (*Cont'd*)

Psychometric properties, research questions, and analysis, with statistic	Criterion	Software package	Refs.	PROMIS-PI results (40 items)		PROMIS-PB results (39 items)	
				Unscaled	Scaled	Unscaled	Scaled
Cross-cultural validity‡							
Absence of differential item functioning (DIF) for McFadden's pseudo R2 change, no. of items	–	R, Lordif 0.3-3	49,50	–	–	–	–
Language (Dutch vs. Flemish)	>2%	–	18	0	–	1 (PAINBE16)	–
Measurement invariance§							
Absence of DIF for McFadden's pseudo R2 change, no. of items	–	R, Lordif 0.3-3	(49,50)	–	–	–	–
Age: median split	>2%	–	18	0	–	0	–
Sex: male vs. female	>2%	–	18	0	–	1 (PAINBE27)	–
Administration mode: digital vs. paper	>2%	–	18	0	–	0	–
Disease activity: DAS28 <3.2 vs. ≥3.2	>2%	–	18	0	–	0	–
Construct validity¶							
Correlation of PROMIS T scores with Pearson correlations							
PROMIS global health pain intensity item	>0.50	–	–	0.80	–	0.61	–
HAQ DI	>0.50	–	–	0.71	–	NA	–
SF-36 PF10	>0.50	–	–	−0.71	–	NA	–
Floor and ceiling effects#	–	–	–	Overall	–	–	–
Floor effects**							
PROMIS-PI	–	–	–	0	–	–	–
PROMIS-PB	–	–	–	0	–	–	–
HAQ DI	–	–	–	0	–	–	–
SF-36 PF10	–	–	–	1.4	–	–	–
RAID	–	–	–	0	–	–	–
Ceiling effects††							
PROMIS-PI	–	–	–	0	–	–	–
PROMIS-PB	–	–	–	0	–	–	–
HAQ DI	–	–	–	19.4	–	–	–
SF-36 PF10	–	–	–	8.3	–	–	–
RAID	–	–	–	5.4	–	–	Q

* PROMIS-PI = Patient-Reported Outcomes Measurement Information System pain interference; PROMIS-PB = PROMIS pain behavior; Refs. = references; IRT = item response theory; PAINBE16 AND PAINBE27 = items from item bank; DAS28 = Disease Activity Score in 28 joints; HAQ DI = Health Assessment Questionnaire disability index; SF-36 = Short Form 36 health survey; PF10 = 10-item physical function subscale; RAID = Rheumatoid Arthritis Impact of Disease.
† Do the items meet the assumptions of and fit to the IRT model, and thus, is it valid to calculate IRT model–based estimates of a patient's score?
‡ Are the item parameters equivalent between cultural or language groups, and thus, is it valid to use the same IRT model across cultural or language groups?
§ Are the item parameters equivalent between the comparison groups at issue, and, thus, is it valid to use the same IRT model to compare these groups?
¶ Does the measure really assesses the intended construct, and thus, are hypotheses addressing its relation with other variables or constructs met?
# Does the measure have a substantial number of patients with the worst possible score (floor effect) or best possible score (ceiling effect), and thus, is the measure's ability to detect differences between or changes within groups of patients limited?
** Percentage of patients with poor health status who obtained the worst possible score. For the PROMIS-PI, PROMIS-PB, HAQ DI, and RAID, lower scores indicate low pain interference, pain behavior, disability, or impact, respectively, and therefore good health status. For the SF-36 PF10, higher scores indicate better physical function and therefore good health status.
†† Percentage of patients with very good health status who obtained the best possible score. For the PROMIS-PI, PROMIS-PB, HAQ DI, and RAID, lower scores indicate low pain interference, pain behavior, disability, or impact, respectively, and therefore good health status. For the SF-36 PF10, higher scores indicate better physical function and therefore good health status.

0–10), HAQ DI (range 0–3), and RAID (range 0–10), higher scores indicate more pain intensity, disability, or impact, respectively. For the SF-36 PF10 (range 0–100), lower scores indicate worse physical function. Evidence supporting the validity of the SF-36 PF10, HAQ DI, and RAID is available for patients with RA (28–34).

Statistical analysis. Demographic and clinical characteristics of study participants were given as descriptive statistics. The psychometric analyses were conducted according to the PROMIS analysis plan (18). Table 1 shows a detailed overview of the psychometric properties that were studied, the research questions addressed, the analyses, statistics, criteria, and software packages used.

PROMIS banks have been developed based on IRT methods, and the estimates of the patient scores are based on the underlying IRT model per bank. The IRT model assumptions and fit were evaluated. Cross-cultural validity was studied by examining differential item functioning (DIF) for language (Dutch versus Flemish). The absence of DIF indicates that Dutch and Flemish patients with similar levels of pain interference or pain behavior respond similarly to the items. Measurement invariance was studied by examining DIF for age, sex, administration mode, and disease activity (DAS28 <3.2 [remission or low disease activity] versus DAS28 ≥3.2 [moderate or high disease activity] [35]).

Construct validity was studied by testing the hypothesis that the PROMIS-PI and PROMIS-PB banks would have strong correlations ($r > 0.50$) with the PROMIS global health pain intensity item. Moreover, we hypothesized that the PROMIS-PI bank would have a strong correlation ($r > 0.50$) with the SF-36 PF10 and HAQ DI because we believe that pain interference and disability are closely related constructs in patients with RA, as disability in patients with RA often results from pain (19).

Reliability indicates the precision of the estimated patient scores of a measure (10,11). In the context of IRT, the precision of the scores can differ across the scale, e.g., the precision of the scores can be higher in the center of the scale and lower at the ends. The IRT scores consist of so-called theta scores. For PROMIS instruments, these scores were estimated, according to the current recommendation, using the US item parameters (10,11,36). Subsequently, the theta scores were transformed into T scores, in which 50 represents the average score of the general US population, with an SD of 10.

Plots were drawn for the PROMIS-PI that show measurement precision (expressed as SEs of theta) across the score range of the total bank, the standard 4- and 8-item short forms (version 1.0.4a and version 1.0.8a), and 3 different simulated CATs. In the first 2 simulated CATs, a fixed number of 4 and 8 items was administered to compare the reliability of these CATs with the corresponding short forms. In the third CAT, the standard PROMIS CAT stopping rules were applied: stop when an SE of ≤3 on the T score scale is reached (comparable to a reliability coefficient >0.90, the accuracy to measure individual patients [23,24]), or when 12 items have been administered. Likewise, plots were drawn for the PROMIS-PB, showing the SEs across the score range of the total bank, the standard 7-item short form (version 1.0.7a), a 7-item simulated CAT, and a standard PROMIS CAT. The T score density plot added the relation between the precision of the measures and the distribution of T scores in the RA sample. Ideally, the precision of the measures is high in the range where the patients are located on the T score scale. We used the R package CatR (version 3.12) for all CAT simulations (37). Floor and ceiling effects of the PROMIS banks were evaluated by counting the number (percentage) of patients with the worst and best possible score, respectively, and were compared to the SF-36 PF10, HAQ DI, and RAID.

RESULTS

Study participants. Figure 1 summarizes the patient selection, and Table 2 shows the demographic and clinical characteristics of the PROMIS-PI and PROMIS-PB samples, including their scores on the banks and legacy instruments.

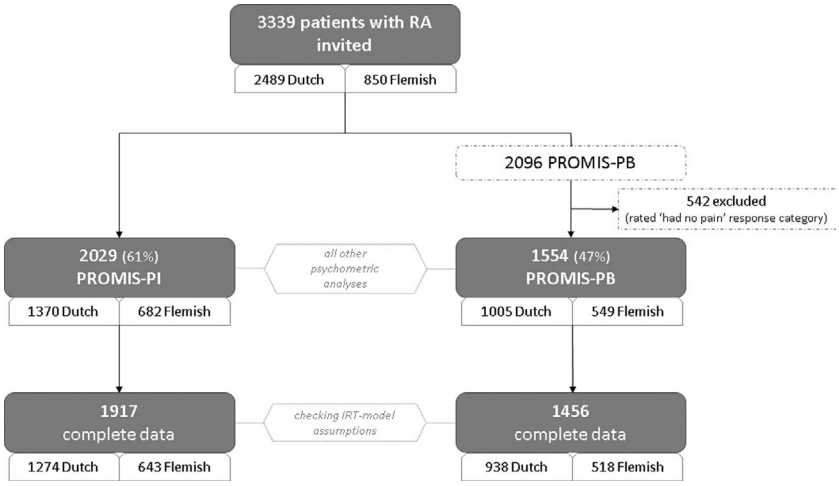

Figure 1. Flowchart of the patient selection. RA = rheumatoid arthritis; PROMIS-PB = Patient-Reported Outcomes Measurement Information System pain behavior; PROMIS-PI = PROMIS pain interference; IRT = item response theory.

Table 2. Demographic and clinical characteristics of the rheumatoid arthritis sample for the Patient-Reported Outcomes Measurement Information System (PROMIS) pain interference (PI) and pain behavior (PB) item banks*

Characteristics	PROMIS-PI (n = 2,029)	PROMIS-PB (n = 1,554)
Age, years		
Mean ± SD	59 ± 13	59 ± 12
Range	19–94	19–94
Sex, %		
Male	31	31
Female	69	69
Country of birth, %		
Netherlands	58	56
Belgium	32	34
Other	10	10
Social status, %†		
Single	22	22
Married or living together	72	74
Other	5	4
Educational level, %†		
Less than high school degree	12	13
High school degree	17	17
Some college	38	38
College or advanced degree	33	32
Employment status, %†		
Full-time	20	18
Part-time	19	17
Unpaid, volunteer, household, or student	14	11
Retired or unemployed	42	38
Social benefits, %†		
Sick listed	7	6
Disability benefit	15	17
Unemployment benefit	3	2
Other	5	4
No social benefit	66	44
Duration of rheumatoid arthritis, %		
1–12 months	2	2
1–2 years	7	6
2–5 years	16	15
>5 years	75	76
Disease Activity Score in 28 joints (range 0–10)		
Mean ± SD	2.6 ± 1.2	2.9 ± 1.2
Range	0–7.5	0–7.5
<3.2 remission or low disease activity, %	77	76
≥3.2 moderate or high disease activity, %	23	24
PROMIS T score‡		
Mean ± SD	53.6 ± 9.9	56.6 ± 5.1
Range	37.4–76.6	38.4–73.4
Generic and disease-specific instruments, mean ± SD		
PROMIS global health pain intensity (range 0–10)	3.6 ± 2.6	4.4 ± 2.3
SF-36 PF10	63.6 ± 27.0	57.3 ± 26.2
HAQ DI	0.9 ± 0.7	1.1 ± 0.7
0 to <1 mild to moderate disability, %	55	45
1 to <2 moderate to severe disability, %	35	42
2–3 severe to very severe disability, %	10	13
RAID	3.1 ± 2.3	3.8 ± 2.1

* For the PROMIS global health pain intensity item, Health Assessment Questionnaire disability index (HAQ DI; range 0–3), and Rheumatoid Arthritis Impact of Disease (RAID; range 0–10), higher scores indicate higher pain intensity, disability or impact, respectively. For the Short Form 36 health survey physical functioning scale (SF-36 PF10; range 0–100), higher scores indicate better physical function.
† Multiple answers were allowed.
‡ PROMIS T score: higher scores indicate more pain interference and pain behavior.

Psychometric properties. *IRT model assumptions and fit.* The IRT model assumptions were met for both banks, and the results showed good fit of the IRT model (Table 1). This result legitimizes the calculation of T scores, based on the IRT model, in patients with RA. Moreover, it legitimizes the use of short forms and CATs because they are also based on the IRT model.

Cross-cultural validity. Dutch and Flemish language groups showed equivalent item parameters (Table 1). None of the PROMIS-PI and only 1 of 39 PROMIS-PB items showed DIF between the Dutch and Flemish languages (PAINBE16 "When I was in pain I appeared upset or sad"; Flemish patients who had similar levels of pain behavior because Dutch patients were slightly more likely to endorse this item). However, the impact of the DIF of this item on the T scores was negligible. This finding indicates that the same IRT model can be used in Dutch as well as Flemish patients, and scores of Dutch and Flemish patients can be compared directly.

Measurement invariance. The DIF analyses were conducted on the total patient group, given the equivalence of the item parameters between Dutch and Flemish language groups. Results showed that the item parameters were equivalent (almost no DIF) in patients differing in age, sex, administration mode, or disease activity (Table 1). None of the items of the PROMIS-PI and only 1 of 39 items of the PROMIS-PB was flagged for DIF, and only for sex (PAINBE27 "I had pain so bad it made me cry"; women who had similar levels of pain behavior because men were slightly more likely to endorse this item). However, the impact of the DIF of this item on the T scores was negligible. This result indicates that the same IRT model can be used within and to compare subgroups of patients who differ with respect to age, sex, administration mode, or disease activity.

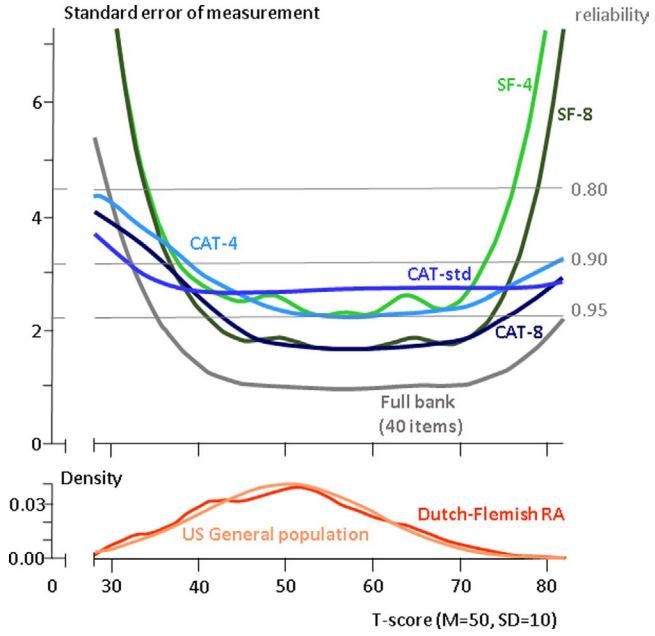

Figure 2. SEs of theta across the range of the Patient-Reported Outcomes Measurement Information System (PROMIS) pain interference T scores. Upper plot shows the total PROMIS pain interference item bank (40-item bank), the standard 4- and 8-item Short Form (SF), the 4- and 8-item simulated computerized adaptive testing (CAT), and the standard PROMIS CAT. The horizontal axis represents the different pain interference levels, with T = 50 representing the mean of the US general population with an SD of 10. The vertical axis represents the SE of theta (reliability), with reference reliabilities of 0.80, 0.90, and 0.95. The lower the curve, the lower the SE, and the greater the reliability. The lower plot shows the distribution of the rheumatoid arthritis (RA) sample along the T score scale.

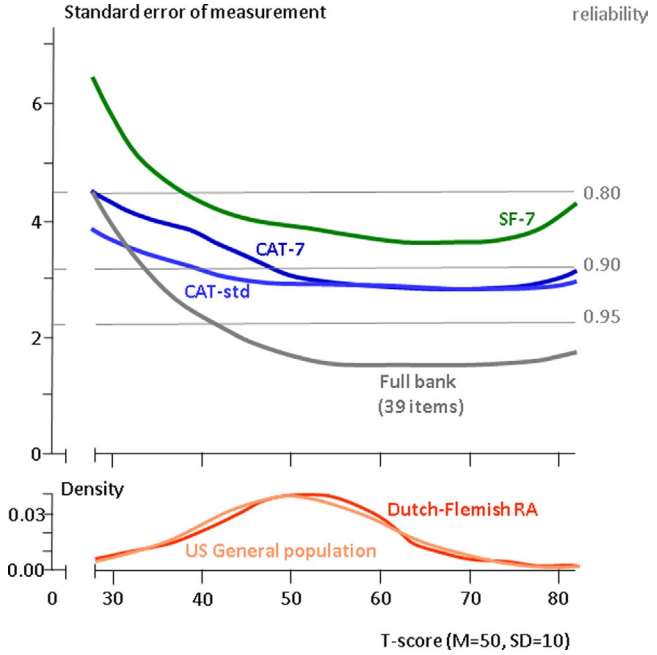

Figure 3. SEs of theta across the range of the Patient-Reported Outcomes Measurement Information System (PROMIS) pain behavior T scores. Upper plot shows the total PROMIS pain behavior item bank (39-item bank), the standard 7-item Short Form (SF), the 7-item simulated computerized adaptive testing (CAT), and the standard PROMIS CAT. The horizontal axis represents the different pain behavior levels, with T = 50 representing the mean of the US general population with an SD of 10. The vertical axis represents the SE of theta (reliability), with reference reliabilities of 0.80, 0.90, and 0.95. The lower the curve, the lower the SE, and the greater the reliability. The lower plot shows the distribution of the rheumatoid arthritis (RA) sample along the T score scale.

Construct validity. The PROMIS-PI correlated, as hypothesized, strongly with the Dutch-Flemish PROMIS global health pain intensity (r = 0.80), the SF-36 PF10 (r = –0.71), and HAQ DI (r = 0.71) (Table 1). The PROMIS-PB also correlated strongly with the Dutch-Flemish PROMIS global health pain intensity (r = 0.61) (Table 1). These results contribute to the evidence that the PROMIS-PI and PROMIS-PB banks are really measuring pain interference and behavior, respectively.

Reliability. Figures 2 and 3 indicate good reliability of the PROMIS-PI and PROMIS-PB. The reliability of the total banks, short forms, and CATs was >0.80 or even >0.90 or >0.95 for the range of the scale where the T scores of the RA sample were located. This result indicates that the PROMIS-PI and PROMIS-PB banks measure pain interference and pain behavior precisely in patients with RA, and even precisely enough to be applied in individual patients. The plot of the PROMIS PI demonstrates that the short forms and corresponding

simulated fixed CATs show equal reliability. In comparison to the 4-item short form and 4-item simulated CAT, the standard PROMIS CAT showed almost equal reliability. However, this finding reflects the fact that the standard PROMIS CAT stops administering items when r > 0.90. The standard PROMIS CAT also needed only a mean number of 4 items (range 2–12) to achieve r, and 66% of the patients needed even <4 items. The plot of the PROMIS-PB shows that, in comparison to the short form, the fixed simulated CAT as well as the standard PROMIS CAT show greater reliability. In comparison to the 7-item simulated CAT, the standard PROMIS CAT shows equal reliability, albeit with a mean of only 5 items (range 2–12), and 70% of the patients needed <7 items.

Floor and ceiling effects. The PROMIS-PI and PROMIS-PB banks showed no floor or ceiling effects (Table 1). The 3 commonly used PROMs within RA had high ceiling (SF-36 PF10, HAQ DI, and RAID) and floor (SF-36 PF10) effects (Table 1).

DISCUSSION

To the best of our knowledge, this is the first extensive study worldwide of the psychometric properties of the PROMIS-PI and PROMIS-PB banks in patients with RA. The results indicate good psychometric properties of both the PROMIS-PI and PROMIS-PB banks, short forms, and CATs in patients with RA. The IRT model assumptions were met, and good IRT model fit was found. Both banks showed good cross-cultural validity, measurement invariance related to age, sex, administration mode, and disease activity, good construct validity, high precision in the range of the patients with RA, and an absence of floor and ceiling effects.

A study strength is that RA samples with a broad spectrum of disease activity (remission to high disease activity) and disability (mild to very severe disability) were included. Despite this broad spectrum, the distribution and mean of the PROMIS-PI and PROMIS-PB T scores of the RA samples were only slightly different from the US general population. This finding suggests that most patients with RA are currently doing quite well. The measurement invariance related to disease activity (no DIF for disease activity) provides evidence that both item banks, even though most participating patients (77%) were in remission or had low disease activity, can also be applied in patients with moderate or high disease activity.

Both the PROMIS-PI and PROMIS-PB banks have already been applied as short forms and CATs in patients with RA (20), and our study provides evidence that it is legitimate to do so. Pain, pain interference, and pain behavior are important constructs to measure in RA (5,7), which is supported by the recent literature showing that a subgroup of early RA patients reports remaining pain that is not inflammatory mediated (38).

Our study is the first to analyze cross-cultural validity of PROMIS instruments between Dutch and Flemish language groups. We found equivalent item parameters between Dutch and Flemish language groups, which indicates that the same IRT models for the PROMIS-PI and PROMIS-PB can be used across different language and disease groups, and that scores between Dutch and Flemish patients can be compared directly.

The results of the current study, addressing the Dutch-Flemish PROMIS-PI and PROMIS-PB banks, can most likely be generalized to the original American-English PROMIS-PI and PROMIS-PB banks. We previously examined DIF (Dutch/Flemish versus English) in chronic pain patients and found no DIF (25,27). In another study, we examined DIF between patients with chronic pain and RA and also found no DIF (39). Therefore, we think that the probability that these items show language DIF in patients with RA is very low. As a consequence, our study results regarding the good psychometric properties of the PROMIS-PI and PROMIS-PB banks in Dutch- or Flemish-speaking patients with RA, combined with the previous studies' results showing an absence of DIF between Dutch- and English-speaking patients, additionally legitimize the use of the PROMIS-PI and PROMIS-PB banks in American English-speaking patients with RA.

Recently, Nagaraja et al (40) identified severity cut points for the PROMIS-PI T scores from the perspectives of patients with RA and their clinical providers. From the perspective of patients, PROMIS-PI T scores of <50 correspond with "no," 50–60 with "mild," 60–65 with "moderate," and >65 with "severe" problems. From the perspective of the clinical providers, the corresponding T score classifications were <50, 50–60, 60–70, and >70, respectively. These results add clinically meaningful interpretations to the PROMIS T scores.

In addition to our study, studies by Khanna et al (4), Orbai and Bingham (41), Bartlett et al (19), Bingham et al (20), and Witter (14) all address the advantages and opportunities for the use of PROMIS short forms and CATs in patients with RA. These studies have shown, among others, that PROMIS CATs are more applicable for use in daily clinical practice than traditional PROMs because CATs are tailored to the individual's ability and are more efficient and precise than other PROMs. For instance, Rose et al (42) showed that the PROMIS physical function CAT was more precise than the HAQ DI (42). The study of Bingham et al (20) is an illustrative example of applying PROMIS instruments in clinical RA practice, outlining practical elements of implementing PROMIS short forms and CATs. They reported that patients with RA specified a preference for CAT over traditional paper forms (20). All studies mentioned address the fact that PROMIS facilitates disease progress monitoring, patient-provider communication, and shared decision-making. Several so-called "cross-walk" tables are available and can be used to convert scores of traditionally used PROMs into PROMIS T scores, e.g., the HAQ DI scores into PROMIS physical function T scores (43). With the use of these tables, historical data can be mapped onto the PROMIS T score scale, enabling a switch from traditional PROMs to PROMIS measures with preservation of historical data.

PROMIS instruments are based on well-developed conceptual models, have been developed by extensive qualitative

research with patients, and have been developed for elementary school reading levels (4,17). Overall, PROMIS instruments are less burdensome and are more precise than traditional PROMs and are easy to interpret. Although PROMIS instruments have advantages over traditional PROMs, and this study has provided evidence of the good psychometric properties of the PROMIS-PI and PROMIS-PB banks in patients with RA, future studies should compare the psychometric properties of the PROMIS-PI and PROMIS-PB instruments directly with other PROMs used for the measurement of pain, such as the visual analog scale and SF-36 bodily pain that are recommended for use in clinical practice in the ICHOM's standard set for inflammatory arthritis, to further reduce the variety of instruments being used in clinical practice (7).

In the Netherlands and Flanders, PROMIS CATs are available through the Dutch-Flemish Assessment Center (www.dutchflemishpromis.nl). As a starting point, US item parameters are used because US item parameters are the current PROMIS recommendation across the world if there is no problematic DIF for language. Using the same item parameters across the world enables international comparisons of PROMIS T scores. More research is needed to identify conditions in which country-specific item parameters may be more valid. In addition, we recommend future studies on the psychometric properties of the PROMIS fatigue item bank in patients with RA, as fatigue, in addition to physical function and pain interference, is also a core outcome in RA.

In conclusion, both the PROMIS-PI and PROMIS-PB banks showed good psychometric properties in patients with RA. The highly efficient PROMIS-PI and PROMIS-PB CATs in research and clinical practice are user friendly and feasible, with little administration time, and have the potential for valid and precise standardized and routine patient monitoring of pain interference and pain behavior.

ACKNOWLEDGMENTS

The authors thank the Dutch-Flemish PROMIS group, which is an initiative that aims to translate and implement PROMIS item banks and CATS in the Netherlands and Flanders (www.dutchflemishpromis.nl/). We also thank Jacqueline Bruinsma for all her administrative support and Oguzhan Ogreden for his support with the reliability analyses.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Crins had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Crins, Terwee, Westhovens, van Schaardenburg, Smits, Joly, Verschueren, Van der Elst, Dekker, Boers, Roorda.

Acquisition of data. Crins, Westhovens, van Schaardenburg, Joly, Verschueren, Van der Elst, Boers.

Analysis and interpretation of data. Crins, Terwee, Smits, Roorda.

REFERENCES

1. Hootman JM, Helmick CG. Projections of US prevalence of arthritis and associated activity limitations. Arthritis Rheum 2006;54:226–9.
2. Lawrence RC, Felson DT, Helmick CG, Arnold LM, Choi H, Deyo RA, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part II. Arthritis Rheum 2008;58:26–35.
3. Helmick CG, Felson DT, Lawrence RC, Gabriel S, Hirsch R, Kwoh CK, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. Arthritis Rheum 2008;58:15–25.
4. Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, Hays RD. The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). Arthritis Care Res (Hoboken) 2011;63 Suppl 1:S486–90.
5. Boers M, Tugwell P, Felson DT, van Riel PL, Kirwan JR, Edmonds JP, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol Suppl 1994;41:86–9.
6. Aletaha D, Landewe R, Karonitsch T, Bathon J, Boers M, Bombardier C, et al. Reporting disease activity in clinical trials of patients with rheumatoid arthritis: EULAR/ACR collaborative recommendations. Ann Rheum Dis 2008;67:1360–4.
7. ICHOM. Inflammatory arthritis: the standard set. 2018. URL: http://www.ichom.org/medical-conditions/inflammatory-arthritis/.
8. Bacalao EJ, Greene GJ, Beaumont JL, Eisenstein A, Muftic A, Mandelin AM, et al. Standardizing and personalizing the treat to target (T2T) approach for rheumatoid arthritis using the Patient-Reported Outcomes Measurement Information System (PROMIS): baseline findings on patient-centered treatment priorities. Clin Rheumatol 2017;36:1729–36.
9. Keefe F, Williams D, Smith S. Assessment of pain behaviors. In: Turk D, Melzack R, eds. Handbook of pain assessment. New York (NY): Guilford Press; 2001. p. 170–87.
10. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain 2010;150:173–82.
11. Revicki DA, Chen WH, Harnam N, Cook KF, Amtmann D, Callahan LF, et al. Development and psychometric analysis of the PROMIS pain behavior item bank. Pain 2009;146:158–69.
12. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. BMJ Qual Saf 2014;23:508–18.
13. Porter I, Goncalves-Bradley D, Ricci-Cabello I, Gibbons C, Gangannagaripalli J, Fitzpatrick R, et al. Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. J Comp Eff Res 2016;5:507–19.
14. Witter J. The promise of patient-reported outcomes measurement information system: turning theory into reality. A uniform approach to patient-reported outcomes across rheumatic diseases. Rheum Dis Clin North Am 2016;42:377–94.
15. Gershon R. Computer adaptive testing. J Appl Meas 2005;6:109–27.

16. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol 2010;63:1179–94.
17. Cella D, Gershon R, Lai J, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007;16:133–41.
18. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45:22–31.
19. Bartlett SJ, Orbai AM, Duncan T, DeLeon E, Ruffing V, Clegg-Smith K, et al. Reliability and validity of selected PROMIS measures in people with rheumatoid arthritis. PLoS One 2015;10:e0138543.
20. Bingham CO III, Bartlett SJ, Merkel PA, Thelma J, Pilkonis PA, Edmundson L, et al. Using patient reported outcomes and PROMIS in research and clinical applications: experiences from the PCORI pilot projects. Qual Life Res 2016;25:2109–16.
21. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO III, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569–81.
22. Terwee CB, Roorda LD, de Vet HC, Dekker J, Westhovens R, van Leeuwen J, et al. Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). Qual Life Res 2014;23:1733–41.
23. PROMIS. PROMIS Pain interference Scoring Manual. 2017:1–16. URL: https://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Pain_Interference_Scoring_Manual.pdf.
24. PROMIS. Pain behavior. URL: http://www.healthmeasures.net/images/promis/manuals/PROMIS_Pain_Behavior_Scoring_Manual.pdf.
25. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration of the Dutch-Flemish PROMIS pain behavior item bank in patients with chronic pain. Eur J Pain 2016;20:284–96.
26. Cook KF, Keefe F, Jensen MP, Roddey TS, Callahan LF, Revicki D, et al. Development and validation of a new self-report measure of pain behaviors. Pain 2013;154:2867–76.
27. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration and validation of the Dutch-Flemish PROMIS pain interference item bank in patients with chronic pain. PLoS One 2015;10:1–18.
28. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009;18:873–80.
29. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
30. Ten Klooster PM, Vonkeman HE, Taal E, Siemons L, Hendriks L, de Jong AJ, et al. Performance of the Dutch SF-36 version 2 as a measure of health-related quality of life in patients with rheumatoid arthritis. Health Qual Life Outcomes 2013;11:77.
31. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
32. Fries J. The Health Assessment Questionnaire (HAQ) and the improved HAQ. Clin Exp Rheumatol 2005. 23 Suppl 39:S14–8.
33. Gossec L, Dougados M, Rincheval N, Balanescu A, Boumpas DT, Canadelo S, et al. Elaboration of the preliminary Rheumatoid Arthritis Impact of Disease (RAID) score: a EULAR initiative. Ann Rheum Dis 2009;68:1680–5.
34. Gossec L, Paternotte S, Aanerud GJ, Balanescu A, Boumpas DT, Carmona L, et al. Finalisation and validation of the rheumatoid arthritis impact of disease score, a patient-derived composite measure of impact of rheumatoid arthritis: a EULAR initiative. Ann Rheum Dis 2011;70:935–42.
35. DAS28. DAS28 score. 2017. URL: http://www.das-score.nl/das28/nl/uitleg-das28/de-das28-score.html.
36. Healthmeasures. URL: www.healthmeasures.net.
37. Magis D, Barrada JR. Computerized adaptive testing with R: recent updates of the package catR. J Stat Softw 2017;76:1–19.
38. Altawil R, Saevarsdottir S, Wedrén S, Alfredsson L, Klareskog L, Lampa J. Remaining pain in early rheumatoid arthritis patients treated with methotrexate. Arthritis Care Res (Hoboken) 2016;68:1061–8.
39. Crins MH, Terwee CB, Ogreden O, Schuller W, Dekker P, Flens G, et al. 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. Qual Life Res 2019;28:1231–43.
40. Nagaraja V, Mara C, Khanna PP, Namas R, Young A, Fox DA, et al. Establishing clinical severity for PROMIS measures in adult patients with rheumatic diseases. Qual Life Res 2018;27:755–64.
41. Orbai AM, Bingham CO III. Patient reported outcomes in rheumatoid arthritis clinical trials. Curr Rheumatol Rep 2015;17:28.
42. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE. The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 2014;67:516–26.
43. Choi SW, Podrabsky T, McKinney N, Schalet BD, Cook KF, Cella D. PROsetta Stone methodology: a Rosetta stone for patient reported outcomes. 2015. URL: http://www.prosettastone.org/Methodology/Documents/PROSetta Methodology Report.pdf.
44. Lavaan RY. An R package for structural equation modeling. J Stat Softw 2012;48:1–36.
45. Bentler PM. Comparative fit indexes in structural models. Psychol Bull 1990;107:238–46.
46. Rodriguez A, Reise S, Haviland M. Applying bsfactor statistical indices in the evaluation of psychological measures. J Pers Assess 2016;98:223–37.
47. Van der Ark L. Mokken scale analysis in R. J Stat Softw 2007;20:1–19.
48. Chalmers RP. The mirt package: multidimensional item response theory. Library of the R package. 2016. URL: https://www.jstatsoft.org/article/view/v048i06.
49. Choi SW, Gibbons LE, Crane PK. Package 'lordif': logistic ordinal regression differential item functioning using IRT, version 0.3-3. 2016. URL: https://cran.r-project.org/web/packages/lordif/lordif.pdf.
50. 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.