PROMIS Physical Function short forms display item- and scale-level characteristics at least as good as the Roland Morris Disability Questionnaire in patients with chronic low back pain

Chiarotto, Alessandro; Roorda, Leo D; Crins, Martine H; Boers, Maarten; Ostelo, Raymond W; Terwee, Caroline B

published in
Archives of Physical Medicine and Rehabilitation
2020

DOI (link to publisher)
10.1016/j.apmr.2019.09.018

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

ACRM
AMERICAN CONGRESS OF REHABILITATION MEDICINE

Archives of Physical Medicine and Rehabilitation

journal homepage: www.archives-pmr.org

Archives of Physical Medicine and Rehabilitation 2020;101:297-308

ORIGINAL RESEARCH

PROMIS Physical Function Short Forms Display Item- and Scale-Level Characteristics at Least as Good as the Roland Morris Disability Questionnaire in Patients With Chronic Low Back Pain

Alessandro Chiarotto, PhD,[a,b] Leo D. Roorda, PhD,[c] Martine H. Crins, MSc,[c] Maarten Boers, PhD,[d,e] Raymond W. Ostelo, PhD,[a,b] Caroline B. Terwee, PhD[a]

From the [a]Department of Health Sciences, Faculty of Science, Amsterdam Movement Sciences Research Institute, VU University, Amsterdam; [b]Department of General Practice, Erasmus MC, University Medical Center, Rotterdam; [c]Amsterdam Rehabilitation Research Center, Reade, Amsterdam; [d]Amsterdam Rheumatology and Immunology Center, VU University Medical Center, Amsterdam; and [e]Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Amsterdam UMC, VU University Medical Center, Amsterdam, The Netherlands.

Abstract

Objective: To compare dimensionality, item-level characteristics, scale-level reliability, and construct validity of PROMIS Physical Function short forms (PROMIS-PF) and 24-item Roland Morris Disability Questionnaire (RMDQ-24) in patients with chronic low back pain (LBP).

Design: Cross-sectional study.

Setting: Secondary care center for rehabilitation and rheumatology.

Participants: Patients with nonspecific LBP ≥3 months (N=768). Mean age was 49±13 years, 77% were female, and 54% displayed pain for more than 5 years.

Interventions: Not applicable.

Main Outcome Measures: Dutch versions of the 4-, 6-, 8-, 10-, and 20-item PROMIS-PF and of the RMDQ-24.

Results: PROMIS-PF-6, PROMIS-PF-8, and RMDQ-24 exhibited sufficient unidimensionality (confirmatory factor analysis: comparative fit index>0.950, Tucker-Lewis index>0.950, root means square error of approximation<0.060), whereas the other instruments did not. All instruments were free of local dependence except PROMIS-PF-20 with 4 item pairs with clear residual correlations. Mokken scale analysis found 1 nonmonotone item for PROMIS-PF-20 and 8 for RMDQ-24 (ie, the probability of endorsing these items was not increasing with increasing level on the underlying construct). PROMIS-PF-20 displayed 2 misfitting items (S-χ^2 *P* value>.001). Two-parameter item response theory models found 2 items with low discrimination for RMDQ-24. All other instruments had adequate fit statistics and item parameters. PROMIS-PF-20 displayed the best scale-level reliability. Construct validity was sufficient for all instruments as all hypotheses on expected correlations with other instruments and differences between relevant subgroups were met.

Conclusions: PROMIS-PF-6, PROMIS-PF-8, and RMDQ-24 exhibited better unidimensionality, whereas PROMIS-PF-4, PROMIS-PF-6, PROMIS-PF-8, and PROMIS-PF-10 showed superior item-level characteristics. PROMIS-PF-20 was the instrument with the best scale-level reliability. This study warrants assessment of other measurement properties of PROMIS-PF short forms in comparison with disease-specific physical functioning instruments in LBP.

Archives of Physical Medicine and Rehabilitation 2020;101:297-308

Low back pain (LBP) is globally the most burdensome health condition in terms of disability.[1] There is some agreement that physical functioning (referring to daily physical activities ranging from self-care to more complex activities) is the most

Presented as a poster to the International Back and Neck Pain Research Forum, September 12-15, 2017, Oslo, Norway.

Disclosures: none.

0003-9993/19/$36 - see front matter © 2019 by the American Congress of Rehabilitation Medicine
https://doi.org/10.1016/j.apmr.2019.09.018

important health domain to be measured in LBP.[2] Patient-reported outcome measures (PROMs) are the most frequently used instruments to measure it,[3] and this type of instrument has been advocated to monitor patient outcomes in routine clinical practice and to measure (cost-) effectiveness of interventions.[4,5]

The Roland Morris Disability Questionnaire is the most frequently used instrument to measure physical functioning in LBP clinical trials[6] and has been recommended to measure this domain in clinical research and practice.[7-10] Various versions exist,[11] but the original 24-item Roland Morris Disability Questionnaire (RMDQ-24) is the most frequently used.[12] The RMDQ-24 was developed by selecting items from the Sickness Impact Profile that could reflect physical functions that were likely to be affected by LBP.[13] Twenty of its items measure activity limitations, 2 measure psychological impairments, 1 measures sleep disturbances, and 1 measures pain and symptoms.[14] The measurement properties of the RMDQ-24 have been thoroughly investigated in patients with LBP and have exhibited sufficient test-retest reliability, construct validity, and responsiveness.[15,16] Nevertheless, several studies have found that its structural validity is not sufficient (ie, its total score is not unidimensional) and that its measurement error is too large to distinguish error from "real" change[15-17]; a few recent studies have also highlighted issues with its content validity in measuring physical functioning aspects important to patients with LBP.[17]

The Patient Reported Outcomes Measurement Information System (PROMIS) initiative has developed a series of patient-reported item banks measuring a broad range of health domains.[18,19] These item banks are domain specific, assumed to be applicable across various health conditions, and designed to be administered through computerized adaptive testing (CAT).[20] A combination of qualitative (eg, cognitive interviews with patients) and quantitative (eg, confirmatory factor analysis, item response theory) methods were used for their development.[21,22] Static short forms, including the best performing items, were extracted from each item bank.[18,23] A PROMIS Physical Function (PROMIS-PF) item bank was developed with the goal of measuring a person's ability to carry out various activities that require physical capability, ranging from self-care to more vigorous activities that require increasing degrees of mobility, strength, or endurance.[24-26] PROMIS-PF was evaluated in the general population and in samples including patients with various conditions, exhibiting adequate structural and construct validity.[27-30] For adults, 5 standard static short forms of 4, 6, 8, 10, and 20 items were developed for PROMIS-PF.[31]

The National Institutes of Health Task Force on research standards for chronic LBP recommended the 29-item PROMIS profile, which consists of a set of short forms measuring different domains and includes the 4-item PROMIS-PF short form (PROMIS-PF-4) as a measure of physical functioning.[32] However, the measurement properties of this and other PROMIS-PF short forms have not been assessed as stand-alone instruments in patients with LBP but only in other musculoskeletal disorders.[33-36] For this reason and for their universal (generic) nature, an international Delphi panel of experts recently considered these instruments not ready to be endorsed as core instruments for LBP.[7] Recent research has highlighted the urgent need for head-to-head comparison studies of PROMs measuring the same domain in LBP to determine if any of these should be preferred over others.[7,8,15,17] The RMDQ-24 was developed as a disease-specific tool to measure LBP-related physical disabilities,[12] while PROMIS-PF was designed to assess physical functioning in general.[37] Despite this difference, it is assumed that both instruments can measure physical functioning in patients with LBP because these are the domain and target population for which they have already been recommended by measurement experts.[7,9,10,32]

Various measurement properties of a PROM can be assessed and compared in head-to-head comparison studies.[38] A panel of clinimetric and psychometric experts determined that content validity (ie, the degree to which the content of a PROM is an adequate reflection of the domain to be measured)[38] is the first property to be evaluated when selecting a PROM.[39] The second property to consider is structural validity (ie, the degree to which the items of a PROM assess the domain to be measured and only this domain).[38] Content validity and structural validity are distinguished by the consensus-based standards for the selection of health measurement instruments (COSMIN) taxonomy from construct validity/hypotheses testing (ie, the degree to which the scores of a PROM are consistent with hypotheses [eg, regarding internal relationships, relationships to scores of other instruments, differences between relevant groups] based on the assumption that the instrument validly measures the construct to be measured).[38] Regarding content and structural validity, there is high-quality evidence showing that the structural validity of the RMDQ-24 is not sufficient, whereas evidence on its content validity is more uncertain.[17] To date, it is unclear if the lack of unidimensionality of the RMDQ-24 is a feature in common with other PROMs to measure physical functioning in LBP[17]; additionally, it is unknown if instruments developed more recently with more advanced psychometric methods (eg, PROMIS-PF) display better structural validity, item-level characteristics, and scale-level reliability than traditional tools like the RMDQ-24.[17] Therefore, the goal of this study was to perform a head-to-head comparison of the PROMIS-PF short forms and the RMDQ-24 in patients with LBP, with emphasis on unidimensionality, item-level characteristics, scale-level reliability, and construct validity.

List of abbreviations:

CAT	**computerized adaptive testing**
CFA	**confirmatory factor analysis**
COSMIN	**consensus-based standards for the selection of health measurement instruments**
IRT	**item response theory**
LBP	**low back pain**
NRS	**numeric rating scale**
PROM	**patient-reported outcome measure**
RMDQ-24	**24-item Roland Morris Disability Questionnaire**
PROMIS	**Patient Reported Outcomes Measurement Information System**
PROMIS-PF	**Patient Reported Outcomes Measurement Information System Physical Function**
PROMIS-PF-4	**4-item PROMIS Physical Function short form**
PROMIS-PF-6	**6-item PROMIS Physical Function short form**
PROMIS-PF-8	**8-item PROMIS Physical Function short form**
PROMIS-PF-10	**10-item PROMIS Physical Function short form**
PROMIS-PF-20	**20-item PROMIS Physical Function short form**
PROMIS-GH-10	**10-item PROMIS Global Health short form**

Methods

Participants

This study is a secondary analysis of a cross-sectional sample of patients with nonspecific chronic pain under treatment in Reade, an outpatient secondary care center for rehabilitation and rheumatology in Amsterdam (Netherlands) between September 2010 and November 2014.[27] Adult patients (21 years and older) with a musculoskeletal pain complaint of at least 3 months and who provided informed consent for participating to research were included in the sample. The local institutional review board approved the study.

For this study, patients with chronic LBP (≥3mo) were selected from the original sample.[27] Patients who had all items missing on the PROMIS-PF short forms or on the RMDQ-24 were excluded. Other patients with missing data were included in the analyses because item response theory (IRT) analysis can handle the presence of missing data.[40] Data were available on sociodemographic and clinical characteristics.

Measurement instruments

The 121-item Dutch version of the PROMIS-PF v1.2 item bank includes items on functioning of the spine, the extremities, and ability to carry out instrumental activities of daily living (eg, housework).[25,26,41] The items of the 4-, 6-, 8-, 10-, and 20-item PROMIS-PF short forms were extracted from this item bank. The 4-, 6-, and 8-item PROMIS-PF (PROMIS-PF-4, PROMIS-PF-6, PROMIS-PF-8) were developed by using quantitative analyses, for example, maximum interval information and CAT simulations, and qualitative analysis, for example, interviewing content experts.[31] The quantitative part was conducted in an internet sample of more than 21,000 persons from the general population, including various clinical samples, among which were 1473 adults with self-reported osteoarthritis or rheumatoid arthritis.[23,25] The items for these 3 forms were selected so that the 3 instruments could be nested (table 1).[31] The 10- and 20-item PROMIS-PF (PROMIS-PF-10, PROMIS-PF-20) were constructed by the PROMIS domain team with a focus on representing the breadth of the measured construct and the content of the item bank[31]; PROMIS-PF-10 is nested in the PROMIS-PF-20 (see table 1). A

Table 1 Content, missing data, and descriptive statistics of the items included in the PROMIS Physical Function short forms in patients with chronic low back pain (N = 768)

		PROMIS-PF					Missing	
Item Code	Item Content	4	6	8	10	20	Responses (n)	Mean ± SD
	Are you able to							
PFA11	do chores, such as vacuuming or yard work?*	✓	✓	✓	✓	✓	3	2.4±1.2
PFA21	go up and down stairs at a normal pace?*	✓	✓	✓			2	2.9±1.3
PFA23	go for a walk of at least 15 minutes?*	✓	✓	✓			1	3.5±1.4
PFA53	run errands and shop?*	✓	✓	✓			5	3.1±1.2
	Does your health now limit you in							
PFC12	doing 2 hours of physical labor?†		✓	✓		✓	7	2.2±1.1
PFB1	doing moderate work around the house like vacuuming, sweeping floors, or carrying in groceries?†		✓	✓			7	2.3±1.1
PFA5	lifting or carrying groceries?†			✓	✓	✓	2	2.4±1.0
PFA4	doing heavy work around the house like scrubbing floors or lifting or moving heavy furniture?†			✓			5	1.6±0.9
PFA1	doing vigorous activities, such as running, lifting heavy objects, or participating in strenuous sports?†				✓	✓	5	1.4±0.8
PFC36	walking more than 1.5 km?†				✓	✓	5	2.4±1.3
PFC37	climbing 1 flight of stairs?†				✓	✓	6	3.3±1.2
PFA3	bending, kneeling, or stooping?†				✓	✓	2	2.5±1.0
	Are you able to							
PFA16	dress yourself, including tying shoelaces and doing buttons?*				✓	✓	4	2.9±1.3
PFB26	shampoo your hair?*				✓	✓	3	4.1±1.1
PFA55	wash and dry your body?*				✓	✓	3	4.0±1.1
PFC45	get on and off the toilet?*				✓	✓	5	3.9±1.1
PFA12	push open a heavy door?*					✓	3	2.6±1.2
PFA34	wash your back?*					✓	4	3.0±1.4
PFA38	dry your back with a towel?*					✓	7	3.9±1.3
PFA51	sit on the edge of a bed?*					✓	4	4.5±0.9
PFA56	get in and out of a car?*					✓	6	3.6±1.1
PFB19	squeeze a new tube of toothpaste?*					✓	4	4.4±1.0
PFB22	hold a plate full of food?*					✓	2	4.1±1.1
PFB24	run a short distance, such as to catch a bus?*					✓	2	2.5±1.4
PFC46	transfer from a bed to a chair and back?*					✓	5	4.2±1.1

* These items can be scored as "without any difficulty," "with a little difficulty," "with some difficulty," "with much difficulty," and "unable to do."
† These items can be scored as "not at all," "very little," "somewhat," "quite a lot," and "cannot do."

Table 2 Content, missing data, and descriptive statistics of the items included in the 24-item Roland Morris Disability Questionnaire in patients with chronic low back pain (N = 768)

Item Code	Item Content	Missing Responses (n)	Item Endorsement (%)
1	I stay at home most of the time because of my back	6	60
2	I change position frequently to try and get my back comfortable	5	94
3	I walk more slowly than usual because of my back	5	74
4	Because of my back pain, I am not doing any of the jobs that I usually do around the house	5	68
5	Because of my back, I use a handrail to get upstairs	13	76
6	Because of my back, I lie down to rest more often	8	62
7	Because of my back, I have to hold on to something to get out of an easy chair	8	60
8	Because of my back, I try to get other people to do things for me	6	44
9	I get dressed more slowly than usual because of my back	8	62
10	I only stand for short periods of time because of my back	10	35
11	Because of my back, I try not to bend or kneel down	9	67
12	I find it difficult to get out of a chair because of my back	5	59
13	My back is painful almost all the time	4	70
14	I find it difficult to turn over in bed because of my back	4	65
15	My appetite is not very good because of my back pain	11	17
16	I have trouble putting on my socks (or stockings) because of the pain in my back	6	64
17	I only walk short distances because of my back	3	67
18	I sleep less well because of my back	4	65
19	Because of my back pain, I get dressed with help from someone else	8	10
20	I sit down for most of the day because of my back	8	33
21	I avoid heavy jobs around the house because of my back	6	81
22	Because of my back pain, I am more irritable and bad tempered with people than usual	9	49
23	Because of my back, I go up stairs more slowly than usual	7	74
24	I stay in bed most of the time because of my back	6	11

time frame is not provided for any item, but current health status is inferred; each item is scored on a 5-point Likert scale. T scores with a mean of 50±10 representing the (Unites States) population were calculated for each short form with the scoring service of the HealthMeasures Assessment Center,[42] with higher scores indicating better functioning.

The RMDQ-24 includes 24 statements representing activities routinely done or avoided and asks respondents to endorse those that describe themselves "today" (ie, dichotomous responses)[12] (table 2). A 0-24 sum score is calculated by counting the number of endorsed items[12] and can be converted into a 0-100 total score,[43] with higher scores indicating worse functioning.

PROMIS-PF short forms and RMDQ-24 were administered digitally with a computer device to every patient in the same measurement occasion, and the PROMIS-PF was administered prior to the RMDQ-24.

Comparator measurement instruments

Comparator instruments to assess construct validity were chosen based on the variables available in the data set and domains frequently measured in LBP. Recent consensus exercises established that, besides physical functioning, pain intensity and health-related quality of life should always be measured for research and practice in patients with LBP.[2,44]

The 10-item PROMIS Global Health v1.2 short form (PROMIS-GH-10) assesses generic domains of health and well-being: self-rated health, quality of life, physical functioning, psychological functioning, (satisfaction with) social functioning, fatigue, and pain.[45] Nine items are scored on a 5-point Likert scale; 1 pain item is scored on a 0-10 numeric rating scale (NRS) ranging from "no pain" to "worst imaginable pain." Two raw summary v1.2 physical and mental health component scores were calculated and converted into T scores, with higher scores indicating better health.[4] The PROMIS-GH-10 v1.2 showed favorable measurement properties in the general population in samples of patients undergoing knee arthroscopy and with fibromyalgia.[35,46,47]

The 11-point NRS pain item of the PROMIS-GH-10 was also used as stand-alone pain intensity instrument because it is recommended for measuring pain intensity in chronic LBP[7,8,32]; this instrument has exhibited sufficient test-retest reliability and construct validity in this patient population.[48]

Statistical analysis

Descriptive statistics were used for sociodemographics, clinical characteristics, and calculating instruments' total scores. IRT analyses were also used in this study because IRT provides an excellent toolbox for psychometric evaluations by focusing on the relationship between item responses and a respondent's level on the underlying measured domain.[49,50] Some analytic features of IRT cannot be obtained with classical test theory analysis, such as estimating item parameters and scale-level reliability along the continuum representing the construct and examining the optimal number of response options for each item.[49,50] The PROMIS IRT analysis plan was followed to assess IRT assumptions and model fit.[22] In addition, construct validity was assessed.

IRT assumptions

Unidimensionality indicates that all the items of a questionnaire measure only 1 single underlying construct; it legitimizes that the item scores can be used to calculate total score. Unidimensionality was evaluated with a confirmatory factor analysis (CFA) on the polychoric correlation matrix through application of a diagonally weighted least squares estimator.[22] Fit to a unidimensional model was considered sufficient if the CFA scaled parameters met the following indices: comparative fit index>0.950, Tucker-Lewis index>0.950, and root means square error of approximation<0.060.[39] CFA was evaluated with the R package *lavaan*.[51]

Local independence means that responses to the items are independent of each other after controlling for the dominant construct. So, it indicates that the item scores vary only based on the "level" of the construct being measured, in this case physical functioning. Local independence was examined by checking the residual correlation matrix resulting from CFA. The residual correlations are the correlations between the error terms of the items. Item pairs with residual correlations>0.20[22] were considered potentially locally dependent and further examined; an item pair was considered locally dependent if the removal of 1 of the 2 items led to substantial changes in IRT item parameters and fit. An instrument was considered free of local dependence if no locally dependent item pairs could be retrieved.[27] Residual correlations were calculated with the R package *lavaan*.[51]

Monotonicity indicates that the probability of affirmative responses to the items increases with increasing levels on the underlying construct. Lack of monotonicity will result in items with "disordered" response options. Monotonicity was assessed by fitting the nonparametric monotone homogeneity model from Mokken scale analysis, which assesses whether a cluster of items adheres to this measurement model.[52,53] Scalability coefficients were calculated for each item (H_i), expressing the degree to which an item is related to other items in the scale.[54] The resulting item response curves were inspected to determine the presence of nonmonotone items; an instrument was considered free of monotonicity if it did not include any nonmonotone item.[22] The R package *mokken* was used for this analysis.[55]

IRT model fit and item parameters

IRT model fit indicates that the responses to the items can be described sufficiently by the IRT model at issue and is a prerequisite for calculating IRT parameters (ie, the so-called item slopes and item thresholds) and IRT-based total scores. To assess IRT model fit, the graded response model for polytomous data[56] was used for PROMIS-PF short forms, and the 2-parameter logistic model was used for dichotomous data[57] for the RMDQ-24. Model fit was assessed with S-χ^2 item fit statistics, which quantify differences between observed and expected response frequencies under the estimated IRT model, with S-χ^2 P value<.001 indicating item misfit.[22] An instrument was evaluated free of model misfit if it did not include any misfitting item.[22] IRT analyses were undertaken with the R package *mirt*.[58]

Item slopes (α) indicate how discriminative an item is in measuring the underlying construct. The higher α, the higher the discrimination of the item at issue and the higher its ability to distinguish between respondents (patients) who only have a small difference in level on the construct (called theta (θ) in IRT, in this case physical functioning. Item thresholds (β), which are the thresholds between the response options of each item, are indicators of their "level" on the construct and are located along theta. Item thresholds should have increasing values on theta and not be disordered.[49] An item was considered to be sufficiently discriminative if its item slope was >1. Item thresholds were considered to be sufficient if they were ordered along theta as expected. When removing a potentially locally dependent item, a change in IRT fit and parameters was considered substantial if 1 or more misfitting items displayed adequate fit, 1 or more item slopes shifted from <1 to >1, or 1 or more disordered item thresholds became ordered.

Reliability

In the context of IRT, reliability is operationalized as the degree of information that it provides on the measured construct. The higher the information, the less error, the more precise the construct at issue is being measured, and thus the better the reliability of the instrument. The standard error of measurement of an item or of a questionnaire can be calculated with the formula $SE(\theta) = 1/\sqrt{(information\ (\theta)}$. A smaller SE indicates greater measurement precision, and a 0-1 reliability coefficient can be calculated as 1−(1/information).[59] A scale-level reliability≥0.80 is considered sufficient to analyze population means, while a reliability≥0.90 is required for individual use.[60] Because there is not a standard criterion to judge sufficient reliability within a IRT context, a criterion was specifically made for this study: an instrument was judged to have sufficient reliability if displaying reliability≥0.90 between theta values −4 and 4. A scale information function was estimated for each instrument with the R package *ltm*[61] to have an indication of the measurement precision of the total scale for different values of theta.[22]

Construct validity - hypotheses testing

Construct validity indicates whether an instrument really measures the intended constructs, in this study physical function. If so, hypotheses addressing the relationship between the construct at issue and other variables or constructs should be met. Construct validity was assessed by formulating and testing hypotheses, as suggested by the COSMIN and the International Society for Quality of Life Research initiatives.[62,63]

Based on previous studies in patients with LBP,[15] it was a priori hypothesized that physical functioning instruments would correlate ≥0.60 with the PROMIS-GH-10 physical health score, between 0.20 and 0.50 with the PROMIS-GH-10 mental health score, and between 0.40 and 0.60 with the pain NRS. Pearson correlation coefficients (r) were used to test these hypotheses. It was also hypothesized that patients with concomitant chronic widespread pain would display lower physical functioning than patients without pain because there is evidence that this condition is associated with various comorbidities.[64] To have sufficient construct validity, an instrument had to meet at least 75% of these hypotheses.[39]

Results

Table 3 presents patients' characteristics (N=768). Item-level missing and descriptive responses are outlined in tables 1 and 2.

Table 3 Sociodemographic and clinical characteristics of the patients with chronic low back pain included in this study (N = 768)

Characteristics	Values
Age, mean ± SD (y)	49±13
Female (%)	77
Country of birth (%)	
Netherlands	49
Morocco	9
Turkey	7
Suriname	6
Other countries	12
Missing	17
Education (%)	
No education	5
Primary school	10
Secondary school	36
Higher education	31
Missing	18
Living status (%)	
Single	35
Married living together	53
Living apart together	4
With parents	2
Other situations	5
Missing	1
Work status (%)	
Student	2
Full-time employed	14
Part-time employed	22
Unpaid housekeeping	11
Retired	8
Unemployed	17
Other status	24
Missing	2
When pain started (%)	
3-6 mo ago	1
6 mo-1 y ago	3
1-2 y ago	12
2-5 y ago	29
More than 5 y ago	54
Headache (%)	40
Cancer-related pain (%)	2
Osteoarthritis (%)	41
Rheumatoid arthritis (%)	14
Neuropathic pain (%)	25
Fibromyalgia (%)	37
Neck pain (%)	29
Shoulder pain (%)	34
Chronic widespread pain (%)	56
Other pain disorders (%)	55
PROMIS-PF short forms (0-100), mean ± SD	
4-item	37±7
6-item	36±6
8-item	35±6
10-item	35±6
20-item	35±7
RMDQ-24 (0-100), mean ± SD	56±25
PROMIS-GH-10 short form (0-100), mean ± SD	
Physical health	35±7
Mental health	39±8
Pain NRS (0-10), mean ± SD	6.9±1.9

(*continued on next column*)

IRT assumptions

PROMIS-PF-6, PROMIS-PF-8, and RMDQ-24 exhibited sufficient fit according to CFA (table 4). The other 3 instruments did not meet this fit based on the root means square error of approximation values, which were >0.060 (see table 4), showing less adequate unidimensionality.

No potentially locally dependent item pairs were observed for PROMIS-PF-4, PROMIS-PF-6, and PROMIS-PF-8 (table 5). One potentially locally dependent item pair was found for PROMIS-PF-10, 9 were found for PROMIS-PF-20, and 2 were found for RMDQ-24. For PROMIS-PF-10 and RMDQ-24, item removals did not lead to any substantial change in IRT item parameters. PROMIS-PF-20 presented 4 item pairs with high residual correlations (therefore potentially locally dependent): PFB19-PFB22, PFC12-PFC36, PFC12-PFA1, and PFC36-PFA1 (see table 5). The removal of PFB19, PFB22, PFC12, PFA1, or PFC36 led to an improvement in model fit (ie, no more misfitting items), suggesting that all item pairs were locally dependent.

PROMIS-PF-4, PROMIS-PF-6, PROMIS-PF-8, and PROMIS-PF-10 had only monotone increasing item response curves (see table 5). The PROMIS-PF-20 had 1 item and RMDQ-24 had 8 items with a nonmonotone increasing item response curve (see table 5).

IRT analyses

The PROMIS-PF-20 was the only instrument with misfitting items (PFA38 and PFA55, S-χ^2 P value<.991) (see table 5). All PROMIS-PF short forms had adequate item parameters because all item slopes were ≥1 and all item thresholds were ordered (see table 5). Item thresholds theta values ranged from −1.7 to 1.9 for PROMIS-PF-4, from −1.7 to 2.1 for PROMIS-PF-6, from −1.7 to

Table 4 Results of confirmatory factor analysis on the PROMIS Physical Function short forms and the 24-item Roland Morris Disability Questionnaire in patients with chronic low back pain (N = 768)

Instruments	CFI	TLI	RMSEA
PROMIS-PF short forms			
4-item	0.996	0.989	0.064
6-item	0.994	0.990	0.056
8-item	0.991	0.987	0.058
10-item	0.975	0.967	0.085
20-item	0.975	0.972	0.078
RMDQ-24	0.971	0.968	0.054

NOTE. CFI≥0.95, TLI≥0.95, and RMSEA≤0.06 represent sufficient fit.[43].

Abbreviations: CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, root means square error of approximation.

Table 5 Local independence, monotonicity, item response theory parameters, and fit statistics of the items included in the PROMIS Physical Function short forms and in the 24-item Roland Morris Disability Questionnaire in patients with chronic low back pain (N = 768)

Item Code and Abbreviated Text*	LID	MON	H_i	α	β1	β2	β3	β4	*P* Value S-χ^2
PROMIS-PF-4									
PFA11 Do vacuuming or yard work	✓	✓	0.63	2.5	−0.9	0.3	1.1	1.9	.613
PFA21 Go up and down the stairs	✓	✓	0.62	2.8	−1.3	−0.2	0.5	1.5	.008
PFA23 Walk 15 minutes	✓	✓	0.64	2.4	−1.5	−0.7	−0.2	0.6	.969
PFA53 Run errands and shop	✓	✓	0.64	2.6	−1.7	−0.4	0.3	1.1	.101
PROMIS-PF-6									
PFA11 Do vacuuming or yard work	✓	✓	0.65	3.0	−0.8	0.2	1.0	1.9	.091
PFA21 Go up and down the stairs	✓	✓	0.59	1.9	−1.4	−0.2	0.5	1.6	.050
PFA23 Walk 15 minutes	✓	✓	0.62	2.1	−1.6	−0.7	−0.2	0.6	.873
PFA53 Run errands and shop	✓	✓	0.64	2.7	−1.7	−0.4	0.3	1.1	.701
PFC12 Do 2 hours of physical labor	✓	✓	0.61	2.2	−0.7	0.5	1.5	2.1	.705
PFB1 Do moderate work around the house	✓	✓	0.62	2.6	−0.9	0.4	1.4	2.1	.110
PROMIS-PF-8									
PFA11 Do vacuuming or yard work	✓	✓	0.64	2.7	−0.8	0.3	1.1	1.9	.495
PFA21 Go up and down the stairs	✓	✓	0.58	1.8	−1.4	−0.2	0.5	1.6	.253
PFA23 Walk 15 minutes	✓	✓	0.63	2.2	−1.5	−0.7	−0.2	0.6	.516
PFA53 Run errands and shop	✓	✓	0.63	2.4	−1.7	−0.5	0.3	1.2	.712
PFC12 Do 2 hours of physical labor	✓	✓	0.62	2.4	−0.6	0.5	1.4	2.0	.354
PFA5 Lifting or carrying groceries	✓	✓	0.64	2.5	−0.1	0.8	1.4	2.0	.897
PFA4 Do heavy work around the house	✓	✓	0.64	2.7	0.2	1.3	2.0	2.6	.364
PFB1 Do moderate work around the house	✓	✓	0.64	2.7	−0.9	0.4	1.4	2.1	.185
PROMIS-PF-10									
PFA11 Do vacuuming or yard work	✓	✓	0.58	2.0	−0.9	0.3	1.2	2.1	.063
PFA55 Go up and down the stairs	✓	✓	0.60	2.6	−2.6	−1.4	−0.7	0.2	.065
PFB26 Shampoo hair	✓	✓	0.57	2.2	−2.4	−1.4	−0.7	0.0	.654
PFC45 Get on and off the toilet	✓	✓	0.61	2.7	−2.6	−1.2	−0.5	0.2	.803
PFA16 Dress yourself	✓	✓	0.63	3.1	−2.1	−1.1	−0.5	0.3	.045
PFA3 Bending, kneeling, or stooping	✓	✓	0.60	2.3	−1.7	0.3	1.3	2.1	.044
PFA5 Lifting or carrying groceries	✓	✓	0.61	3.0	−1.4	0.3	1.4	2.1	.079
PFC36 Walk more than 1.5 km	✓	✓	0.56	1.7	−0.6	0.3	1.1	1.8	.335
PFC37 Climbing 1 flight of stairs	✓	✓	0.58	2.0	−2.0	−0.7	0.2	1.0	.971
PFA1 Do vigorous activities	✓	✓	0.50	1.3	0.9	2.8	3.0	3.8	.761
PROMIS-PF-20									
PFA11 Do vacuuming or yard work	✓	✓	0.57	2.0	−0.9	0.3	1.2	2.1	.776
PFA12 Do 2 hours of physical labor	✓	✓	0.58	2.0	−1.2	0.1	0.9	1.8	.004
PFA34 Wash the back	✓	✓	0.55	1.9	−1.2	−0.4	0.2	1.2	.176
PFA38 Dry the back with a towel	✓	✓	0.58	2.5	−1.8	−1.2	−0.6	0.2	<.001
PFA51 Sit on bed's edge	✓	✓	0.60	2.6	−2.6	−1.8	−1.2	−0.6	.592
PFA55 Wash and dry the body	✓	✓	0.62	3.1	−2.4	−1.3	−0.7	0.2	<.001
PFA56 Get in and out of care	✓	✓	0.59	2.3	−2.5	−1.1	−0.2	0.9	.658
PFB19 Squeeze a toothpaste's tube	✗	✓	0.56	2.2	−2.8	−1.8	−1.2	-0.5	.040
PFB22 Hold a plate full of food	✗	✓	0.54	2.0	−2.5	−1.5	−0.7	0.0	.407
PFB24 Run short distances	✓	✓	0.53	1.6	−0.6	0.2	0.8	1.9	.971
PFB26 Shampoo hair	✓	✓	0.58	2.4	−2.3	−1.3	−0.7	0.0	.898
PFC45 Get on and off the toilet	✓	✓	0.61	2.8	−2.6	−1.2	−0.5	0.2	.345
PFC46 Transfer from bed to chair and back	✓	✓	0.59	2.6	−2.5	−1.2	−0.6	0.2	.008
PFA16 Dress yourself	✓	✓	0.63	3.3	−2.0	−1.1	−0.5	0.3	.015
PFA3 Bending, kneeling, or stooping	✓	✓	0.59	2.1	−1.7	0.2	1.3	2.2	.769
PFA5 Lifting or carrying groceries	✓	✓	0.60	2.2	−1.4	0.3	1.4	2.1	.139
PFC12 Push open a heavy door	✗	✓	0.55	1.7	−0.8	0.5	1.6	2.4	.125
PFC36 Walk more than 1.5 km	✗	✓	0.54	1.6	−0.6	0.3	1.1	1.9	.020
PFC37 Climbing 1 flight of stairs	✓	✓	0.56	2.0	−2.1	−0.7	0.2	1.0	.539
PFA1 Do vigorous activities	✗	✗	0.49	1.2	0.9	2.2	3.1	3.9	.790
RMDQ-24									
1 Stay at home	✓	✓	0.55	2.3	−0.3				.639
2 Change position frequently	✓	✗	0.33	0.8	3.8				.856

(*continued on next page*)

Table 5 (*continued*)

Item Code and Abbreviated Text*	LID	MON	H_i	α	β1	β2	β3	β4	*P* Value S-χ^2
3 Walk more slowly	✓	✗	0.48	2.3	0.8				.025
4 Not doing jobs around the house	✓	✓	0.45	2.1	0.6				.553
5 Use a handrail to get up stairs	✓	✗	0.44	1.8	1.0				.003
6 Lie down to rest more often	✓	✓	0.35	1.3	0.5				.045
7 Hold on to something to get out of a chair	✓	✓	0.43	1.8	0.3				.086
8 Try to get other people to do things	✓	✓	0.43	1.5	−0.2				.879
9 Get dressed more slowly	✓	✓	0.46	2.4	0.4				.373
10 Stand only for short periods	✓	✓	0.53	2.0	−0.5				.654
11 Do not bend or kneel down	✓	✓	0.40	1.7	0.6				.658
12 Find difficult to get out of a chair	✓	✓	0.45	2.1	0.3				.551
13 Back painful almost all the time	✓	✓	0.37	1.4	0.8				.057
14 Difficult to turn over in bed	✓	✗	0.39	1.7	0.6				.042
15 Do not have very good appetite	✓	✓	0.59	1.7	−1.3				.882
16 Have troubles putting on socks	✓	✓	0.42	1.9	0.5				.745
17 Walk only short distances	✓	✓	0.47	2.6	0.5				.382
18 Sleep less well	✓	✗	0.37	1.5	0.6				.226
19 Get dressed with help from others	✓	✗	0.65	1.9	−1.7				.390
20 Sit down most of the day	✓	✓	0.48	1.5	−0.6				.671
21 Avoid heavy jobs	✓	✓	0.47	1.8	1.2				.851
22 Be more irritable and bad tempered	✓	✗	0.31	0.9	−0.0				.249
23 Go upstairs more slowly	✓	✓	0.53	3.1	0.7				.673
24 Stay in bed most of the time	✓	✗	0.65	2.0	−1.6				.191

NOTE. H_i, Mokken item scalability coefficient (range, 0 to 1); α, item slope (range, 0 to $+\infty$); β, item threshold (range, $-\infty$ to $+\infty$); S-χ^2, item fit statistic (range, 0 to $+\infty$).
Abbreviations: LID, local independence; MON, monotonicity.
* Table 1 and 2 provided the full-item content.

2.6 for PROMIS-PF-8, from −2.6 to 3.8 for PROMIS-PF-10, and from −2.8 to 3.9 for PROMIS-PF-20. The RMDQ-24 had 2 items (2 and 22) with an item slope<1 (see table 5).

PROMIS-PF-10, PROMIS-PF-20, and RMDQ-24 displayed reliability≥0.90 between theta values −4 and 4, but the other 3 PROMIS-PF short forms did not meet the reliability criterion at lower physical functioning levels (fig 1). In an absolute sense, PROMIS-PF-20 was the instrument with the best reliability.

Construct validity: hypotheses testing

Correlations of physical functioning instruments with the PROMIS-GH-10 physical health scores were ≥0.60, as expected (table 6). Correlations with PROMIS-GH-10 mental health scores and NRS were also as expected as well as mean scores in patients with or without chronic widespread pain (see table 6). Each instrument met all the hypotheses (100%) and was considered as having sufficient construct validity.

Discussion

In this study, only PROMIS-PF-8, PROMIS-PF-10, and RMDQ-24 met the criteria for unidimensionality in patients with chronic LBP. PROMIS-PF-4, PROMIS-PF-6, PROMIS-PF-8, and PROMIS-PF-10 displayed better item-level characteristics (ie, local independence, monotonicity, item fit statistics) than PROMIS-PF-20 and RMDQ-24. PROMIS-PF-20, followed by PROMIS-PF-10 and RMDQ-24, was the instrument with the best scale-level reliability (see fig 1). According to these results, it is unclear which instrument is best for researchers and clinicians to measure physical functioning in patients with chronic LBP. Future clinimetric studies should assess other measurement properties of these instruments, such as content validity, test-retest reliability, measurement error, and responsiveness.

This study supports the notion that generic instruments developed with a combination of qualitative and advanced quantitative psychometric methods like the PROMIS-PF short forms may perform better or at least as good as disease-specific instruments like the RMDQ-24.[19,21,22] The PROMIS-PF item bank was developed mainly to be administered as a CAT, [23,25,26] and preliminary evidence shows PROMIS-PF CAT simulations to be more reliable than PROMIS-PF short forms.[23,26,27] Thus, PROMIS-PF CAT administrations are expected to perform even better than PROMIS-PF short forms also in LBP and should be included in future studies comparing physical functioning instruments. PROMIS-PF-20 item-level characteristics were less good than those of the other instruments, possibly because of the irrelevance of some of its items for patients with LBP, for example, the locally dependent item pair PFB19-PFB22 (see table 1) that refers to upper extremity activities. More specifically, PROMIS-PF-20 may need refinement, for example, by replacing the 4 locally dependent items with other items included in the item bank. In support of this suggestion, recent analyses have suggested a slight revision of the PROMIS-PF-20 content for other musculoskeletal disorders.[65,66] Another solution could be to sum locally dependent items together into testlets to absorb the local dependence,[67,68] but this would affect the calculation of the instrument's total score and would require clear instructions. At the same time, PROMIS-PF-20 was the most precise instrument (see fig 1), suggesting that the shortest PROMIS-PF forms may have difficulties capturing the low and high ends of physical functioning.

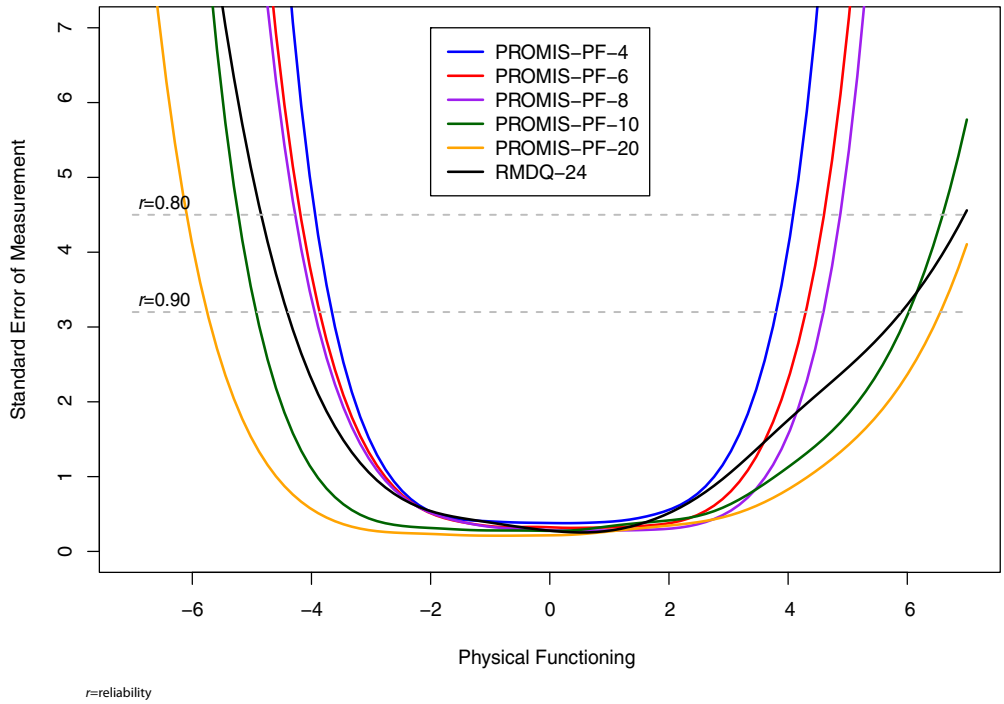

Fig 1 Reliability of the PROMIS Physical Function short forms and the 24-item Roland Morris Disability Questionnaire in measuring different levels (theta) of physical functioning in patients with chronic low back pain (N = 768).

Future studies on psychometric properties should not be limited to comparing structural validity, IRT item-level characteristics and scale reliability, and construct validity. In fact, several psychometric experts feel that content validity is the most important property when selecting a PROM.[39] In a recent systematic review on physical functioning PROMs in LBP, this property was found to be insufficiently investigated and compared across instruments.[17] Other important measurement properties assessed for use in clinical trials or practice are test-retest reliability, measurement error, and responsiveness.[69] It remains unknown if any of the instrument included in this study outperform the others on these properties. Because PROMIS-PF-20 was the instrument with the best measurement precision (see fig 1), larger effect sizes and, thus, responsiveness are to be expected. However, in face of the lack of information on the other important measurement properties, such results should not yet lead to too strict conclusions on which instrument is the best.

This is the first study assessing the PROMIS-PF short forms in a sample including only patients with LBP; therefore, its results cannot be compared with previous studies. Nonetheless, the results of this study highlight that generic short PROMs hold potential to replace longer disease-specific PROMs by displaying similar measurement properties and providing less burden to patients and clinicians. Interestingly, this is the first study finding the RMDQ-24 to be a unidimensional instrument, whereas all previous studies found some departure from unidimensionality.[17] This result may be explained by the CFA estimator used in this study (ie, diagonally weighted least squares), which was not used in previous studies. In previous head-to-head comparison studies, the RMDQ-24 has demonstrated to perform fairly similarly to other broadly used instruments like the Oswestry Disability Index or the Quebec Back

Table 6 Pearson correlation coefficients among comparator instruments and mean scores in relevant subgroups in patients with chronic low back pain (N = 768)

	Correlation Coefficients			Chronic Widespread Pain[§]	
	PROMIS-GH-10				
Instruments	Physical Health*	Mental Health[†]	Pain NRS[‡]	Yes (n = 429)	No (n = 338)
PROMIS PF-4	0.73	0.25	−0.52	36±6	38±7
PF-6	0.74	0.25	−0.55	35±5	37±7
PF-8	0.76	0.24	−0.57	34±5	36±6
PF-10	0.76	0.26	−0.57	34±6	36±7
PF-20	0.76	0.27	−0.58	34±6	36±7
RMDQ-24	−0.61	−0.36	0.55	58±24	54±26

* These correlations were expected to be ≥0.60.
[†] These correlations were expected to be between 0.20 and 0.50.
[‡] These correlations were expected to be between 0.40 and 0.60.
[§] The mean difference of each instrument was expected to be lower (worse physical functioning) in patients with chronic widespread pain.

Pain Disability Scale.[15,70] These previous findings reinforce the need to better understand the relative worth of our study results by simultaneously administering more physical functioning PROMs in a preferably large international population with LBP.

Study limitations

The strengths of this study include the novelty of directly comparing PROMIS-PF short forms with the RMDQ-24, a sufficiently large sample for IRT analysis, and the use of advanced psychometric methods as recommended by the PROMIS and COSMIN initiatives.[22,62] A limitation of this study is the fact that construct validity was assessed only with 2 comparator instruments, and neither measured the same (ie, physical functioning) or a totally unrelated construct. Another potential limitation is that the short forms were extracted from the item bank and not administered as stand-alone instruments. However, it remains unclear if this led to respondent bias because there are no studies showing if the same item administered within different contexts or questionnaires leads to different responses. Another limitation is that, given the heterogeneity of the included sample (see table 3), it cannot be ruled out that the psychometric aspects differed across subgroups of the eligible population; future studies will have to assess for the presence of differential item functioning. Considering the cross-sectional nature of this study, other measurement properties such as test-retest reliability, measurement error, and responsiveness could not be assessed, limiting the applicability of our results.

Conclusions

Item-level characteristics of 4 PROMIS-PF short forms (ie, PROMIS-PF-4, PROMIS-PF-6, PROMIS-PF-8, PROMIS-PF-10) outperformed those of the PROMIS-PF-20 and the RMDQ-24 in patients with chronic LBP. Meanwhile, PROMIS-PF-20 exhibited the best scale-level reliability, followed by PROMIS-PF-10 and RMDQ-24. RMDQ-24 met unidimensionality criteria, whereas PROMIS-PF-4, PROMIS-PF-10, and PROMIS-PF-20 did not. These results suggest that some PROMIS-PF short forms may be preferred over the RDMQ-24 for clinical research and practice in LBP, especially considering that they are shorter instruments, providing less burden to respondents. Additionally, generic instruments like PROMIS-PF short forms may have other advantages over disease-specific PROMs like RMDQ-24 because they can be used to compare patients with different diseases, and they do not require the use of a different PROM for every patient group. However, strong recommendations will require more head-to-head measurement comparison, including other instruments and CAT administrations, and assessing other crucial measurement properties, that is, content validity, test-retest reliability, measurement error, and responsiveness.[39,71]

Keywords

Low back pain; Rehabilitation

Corresponding author

Alessandro Chiarotto, PT, MSc, PhD, Department of Health Sciences, Faculty of Science, Amsterdam Movement Sciences Research Institute, VU University De Boelelaan 1085, Room U-601, 1081HV, Amsterdam, The Netherlands. *E-mail address:* a.chiarotto@vu.nl.

References

1. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016;388:1545-602.
2. Chiarotto A, Deyo RA, Terwee CB, et al. Core outcome domains for clinical trials in non-specific low back pain. Eur Spine J 2015;24:1127-42.
3. Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine 2011;36(21 Suppl):54-68.
4. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346:f167.
5. Coulter A. Measuring what matters to patients. BMJ 2017;356:j816.
6. Froud R, Patel S, Rajendran D, et al. A systematic review of outcome measures use, analytical approaches, reporting methods, and publication volume by year in low back pain trials published between 1980 and 2012: respice, adspice, et prospice. PLoS One 2016;11:e0164573.
7. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in non-specific low back pain. Pain 2018;159:481-95.
8. Chiarotto A, Terwee CB, Ostelo RW. Choosing the right outcome measurement instruments for patients with low back pain. Best Pract Res Clin Rheumatol 2016;30:1003-20.
9. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9-19.
10. Taylor AM, Phillips K, Patel KV, et al. Assessment of physical function and participation in chronic pain clinical trials: IMMPACT/OMERACT recommendations. Pain 2016;157:1836-50.
11. Longo UG, Loppini M, Denaro L, Maffulli N, Denaro V. Rating scales for low back pain. Br Med Bull 2010;94:81-144.
12. Roland M, Morris R. A study of the natural history of back pain: part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8:141-4.
13. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976) 1983;8:141-4.
14. Grotle M, Brox JI, Vøllestad NK. Functional status and disability questionnaires: what do they assess?: a systematic review of back-specific outcome questionnaires. Spine 2005;30:130-40.
15. Chiarotto A, Maxwell LJ, Terwee CB, Wells GA, Tugwell P, Ostelo RW. Roland-Morris Disability Questionnaire and Oswestry Disability Index: which has better measurement properties for measuring physical functioning in nonspecific low back pain? Systematic review and meta-analysis. Phys Ther 2016;96:1620-37.
16. Geere JH, Geere JA, Hunter PR. Meta-analysis identifies back pain questionnaire reliability influenced more by instrument than study design or population. J Clin Epidemiol 2013;66:261-7.
17. Chiarotto A, Ostelo RW, Boers M, Terwee CB. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in low back pain. J Clin Epidemiol 2018;95:73-93.
18. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63:1179-94.
19. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1):3.

20. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007;16:133-41.
21. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45(5 Suppl 1):12.
22. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5 Suppl 1):22-31.
23. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061-6.
24. Bruce B, Fries J, Lingala B, Hussain YN, Krishnan E. Development and assessment of floor and ceiling items for the PROMIS physical function item bank. Arthritis Res Ther 2013;15:R144.
25. Bruce B, Fries JF, Ambrosini D, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11:R191.
26. 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.
27. Crins MHP, Terwee CB, Klausch T, et al. The Dutch-Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol 2017;87:47-58.
28. Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. J Orthop Res 2011;29:947-53.
29. Hung M, Hon SD, Franklin JD, et al. Psychometric properties of the PROMIS physical function item bank in patients with spinal disorders. Spine 2014;39:158-63.
30. Oude Voshaar MA, Peter M, Glas CA, et al. Calibration of the PROMIS Physical Function item bank in Dutch patients with rheumatoid arthritis. PLoS One 2014;9:e92367.
31. A brief guide to the PROMIS Physical Function instruments. Available at: https://assessmentcenter.net/documents/PROMIS%20Physical%20Function%20Scoring%20Manual.pdf. Accessed April 26, 2019.
32. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on research standards for chronic low back pain. J Pain 2014;15:569-85.
33. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the Patient-Reported Outcomes Measurement Information System (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. Ann Rheum Dis 2015;74:104-7.
34. Lee AC, Driban JB, Price LL, Harvey WF, Rodday AM, Wang C. Responsiveness and minimally important differences for 4 Patient-Reported Outcomes Measurement Information System short forms: Physical Function, Pain Interference, Depression, and Anxiety in Knee Osteoarthritis. J Pain 2017;18:1096-110.
35. Merriwether EN, Rakel BA, Zimmerman MB, et al. Reliability and construct validity of the Patient-Reported Outcomes Measurement Information System (PROMIS) instruments in women with fibromyalgia. Pain Med 2016;18:1485-95.
36. Wahl E, Gross A, Chernitskiy V, et al. Validity and responsiveness of a 10-item Patient-Reported Measure of Physical Function in a rheumatoid arthritis clinic population. Arthritis Res Ther 2017;69:338-46.
37. Rose M, Bjorner JB, Becker J, Fries J, Ware J. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2008;61:17-33.
38. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.
39. Prinsen CA, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set"—a practical guideline. Trials 2016;17:449.
40. Mislevy RJ, Wu PK. Missing responses and IRT ability estimation: omits, choice, time limits, and adaptive testing. ETS Research Report Series 1996:1996.
41. Terwee C, Roorda L, De Vet H, 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.
42. PROMIS Instrument Development and Validation Scientific Standards version 2.0. Available at: http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf. Accessed April 26, 2019.
43. Kent P, Lauridsen HH. Managing missing scores on the Roland Morris Disability Questionnaire. Spine 2011;36:1878-84.
44. Clement RC, Welander A, Stowell C, et al. A proposed set of metrics for standardized outcome reporting in the management of low back pain. Acta Orthop 2015;86:523-33.
45. 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.
46. Bryan S, Davis J, Broesch J, et al. Choosing your partner for the PROM: a review of evidence on patient-reported outcome measures for use in primary and community care. Health Policy 2014;10:38-51.
47. Oak SR, Strnad GJ, Bena J, et al. Responsiveness comparison of the EQ-5D, PROMIS Global Health, and VR-12 Questionnaires in knee arthroscopy. Orthop J Sports Med 2016;4. 2325967116674714.
48. Chiarotto A, Maxwell LJ, Ostelo RW, Boers M, Tugwell P, Terwee CB. Measurement properties of Numeric Rating Scale, Visual Analogue Scale and Pain Severity subscale of Brief Pain Inventory in patients with low back pain: a systematic review. J Pain 2019;20:245-63.
49. DeMars C. Item response theory. New York: Oxford University Press; 2010.
50. Embretson SE, Reise SP. Item response theory. Hoboken: Psychology Press; 2013.
51. Rosseel Y. Lavaan: an R package for structural equation modeling and more. Version 0.5—12 (BETA). J Stat Softw 2012;48:1-36.
52. Mokken RJ. A theory and procedure of scale analysis: with applications in political research. Berlin: Walter de Gruyter; 1971.
53. Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. Thousand Oaks: Sage; 2002.
54. Paap MC, Meijer RR, Cohen-Kettenis PT, et al. Why the factorial structure of the SCL-90-R is unstable: comparing patient groups with different levels of psychological distress using Mokken Scale Analysis. Psychiatry Res 2012;200:819-26.
55. van der Ark LA. New developments in Mokken scale analysis in R. J Stat Softw 2012;48:1-27.
56. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika 1970;35:139.
57. Thissen D, Orlando M. Item response theory for items scored in two categories. In: Thissen D, Wainer H, editors. Test scoring. Mahwah: Lawrence Erlbaum Associates; 2001. p 73-140.
58. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Softw 2012;48(6):1-29.
59. Petrillo J, Cano SJ, McLeod LD, Coon CD. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. Value Health 2015;18:25-34.
60. Sijtsma K. Correcting fallacies in validity, reliability, and classification. Int J Test 2009;9:167-94.
61. Rizopoulos D. ltm: an R package for latent variable modeling and item response theory analyses. J Stat Softw 2006;17:1-25.
62. de Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.

63. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 2013;22:1889-905.
64. Kato K, Sullivan PF, Evengård B, Pedersen NL. Chronic widespread pain and its comorbidities: a population-based study. Arch Intern Med 2006;166:1649-54.
65. Beckmann JT, Hung M, Voss MW, Crum AB, Bounsanga J, Tyser AR. Evaluation of the patient-reported outcomes measurement information system upper extremity computer adaptive test. J Hand Surg Am 2016; 41:739-44.
66. Hays RD, Spritzer KL, Amtmann D, et al. Upper-extremity and mobility subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS) adult Physical Functioning item bank. Arch Phys Med Rehabil 2013;94:2291-6.
67. Braeken J. A boundary mixture approach to violations of conditional independence. Psychometrika 2011;76:57-76.
68. Wainer H, Kiely GL. Item clusters and computerized adaptive testing: a case for testlets. J Educ Meas 1987;24:185-201.
69. Wells G, Beaton DE, Tugwell P, et al. Updating the OMERACT filter: discrimination and feasibility. J Rheumatol 2014;41: 1005-10.
70. Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec Back Pain Disability Scale. Measurement properties. Spine 1995;20:341-52.
71. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. J Rheumatol 1998;25:198-9.