Translation, cross-cultural and construct validity of the Dutch–Flemish PROMIS® upper extremity item bank v2.0

Erik-Jan A. Haan[1,2,3] · Caroline B. Terwee[2] · Marieke F. Van Wier[4] · Nienke W. Willigenburg[4] · Derek F. P. Van Deurzen[4] · Martijn F. Pisters[1] · Aaron J. Kaat[5] · Leo D. Roorda[3]

Abstract

Purpose To develop a Dutch–Flemish translation of the PROMIS® upper extremity (PROMIS-UE) item bank v2.0, and to investigate its cross-cultural and construct validity as well as its floor and ceiling effects in patients with musculoskeletal UE disorders.

Methods State of the art translation methodology was used to develop the Dutch–Flemish PROMIS-UE item bank v2.0. The item bank and four legacy instruments were administered to 205 Dutch patients with musculoskeletal UE disorders visiting an orthopedic outpatient clinic. The validity of cross-cultural comparisons between English and Dutch patients was evaluated by studying differential item functioning (DIF) for language (Dutch vs. English) with ordinal logistic regression models and McFadden's pseudo R^2-change of $\geq 2\%$ as critical value. Construct validity was assessed by formulating a priori hypotheses and calculating correlations with legacy instruments. Floor/ceiling effects were evaluated by determining the proportion of patients who achieved the lowest/highest possible raw score.

Results Eight items showed DIF for language, but their impact on the test score was negligible. The item bank correlated, as hypothesized, moderately with the Dutch–Flemish PROMIS pain intensity item (Pearson's $r = -0.43$) and strongly with the Disabilities of the Arm, Shoulder and Hand questionnaire, Subscale Disability/Symptoms (Spearman's $\rho = -0.87$), the Functional Index for Hand Osteoarthritis ($\rho = -0.86$), and the Michigan Hand Outcomes Questionnaire, Subscale Activities of Daily Living ($\rho = 0.87$). No patients achieved the lowest or highest possible raw score.

Conclusions A Dutch–Flemish PROMIS-UE item bank v2.0 has been developed that showed sufficient cross-cultural and construct validity as well as absence of floor and ceiling effects.

Keywords Patient-Reported Outcomes Measurement Information System (PROMIS) · Upper extremity · Cross-cultural validity · Construct validity · Floor and ceiling effects

✉ Leo D. Roorda
l.roorda@reade.nl

1. Physical Therapy Sciences, Program in Clinical Health Sciences, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

2. Department of Epidemiology and Biostatistics, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

3. Amsterdam Rehabilitation Research Center | Reade, Amsterdam, The Netherlands

4. Department of Orthopedic Surgery, Joint Research, OLVG, Amsterdam, The Netherlands

5. Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Introduction

Upper extremity (UE) musculoskeletal disorders are a common health problem, with estimated point prevalence rates ranging from 2 to 53%, with a high burden for patients, health care, and society [1]. With the aging of the population, the burden of this condition is expected to increase further [2]. Patients with UE musculoskeletal disorders suffer from symptoms such as pain and functional decline [3].

Numerous Patient-Reported Outcome Measures (PROMs) for measuring functional status in patients with UE musculoskeletal disorders are used in daily clinical care and in research, but these measures are not without problems [4–9]. There is a lack of convincing evidence regarding their measurement properties [9]. The variety and availability of

multiple PROMs hampers comparability of scores across conditions and settings. Traditional PROMs sometimes contain irrelevant questions, which can lead to incomplete questionnaires and place a high burden on respondents [10, 11]. Thus, several PROMs that are currently used do not meet the recommended minimum standards [12].

The Patient-Reported Outcomes Measurement Information System (PROMIS®) was initiated by six US research institutions and the National Institutes of Health (NIH), with the aim to improve the quality and comparability of health outcome measures, and to reduce the burden for respondents. To achieve this aim, item banks for measuring specified health domains have been developed and validated [13, 14]. An item bank is a set of items (questions), all measuring the same domain, e.g., physical function [15]. The items of an item bank are calibrated on a scale, using Item Response Theory (IRT) modeling, which enables the calculation of precise (reliable) and valid test (total) scores. Moreover, IRT-based item banks enable the use of short forms, i.e., fixed subsets of items from the item bank, and Computerized Adaptive Testing (CAT). CAT uses an algorithm that selects the most informative items from the item bank, based on the individual's responses (answers) to previously administered items. In this way, high precision is combined with low patient burden [16, 17].

PROMIS includes a Physical Function (PROMIS-PF) item bank v2.0, consisting of 165 items, covering central (i.e., spinal), upper, and lower extremity functions, and activities of daily living [18, 19]. Subsets of items from the PROMIS-PF item bank, form an item bank on its own and can be used for measuring lower extremity related (PROMIS Mobility) and UE-related physical function (PROMIS upper extremity [PROMIS-UE]), respectively [20]. Several studies have shown that the precursor of the current PROMIS-UE item bank, v1.2 that included 15 items only, exhibited a ceiling effect [21–25]. The newly developed and extended PROMIS-UE item bank v2.0, which includes 46 items, assesses a wider range of UE functioning which might preclude this ceiling effect [26].

In 2010 the Dutch–Flemish PROMIS group was established, with the aim of translating the PROMIS item banks into Dutch–Flemish and to implement these item banks in the Netherlands and Flanders. Four out of the 46 PROMIS-UE v2.0 items, have not yet been translated into Dutch–Flemish. After translation of the new items, the psychometric properties of the entire Dutch–Flemish PROMIS-UE (DF-PROMIS-UE) item bank v2.0 should be established. Evaluating cross-cultural validity is important in order to determine whether the algorithm, which calculates the IRT-based test scores for American patients, is also applicable for Dutch and Flemish patients. Moreover, this is important to establish the comparability of the scores of US patients versus Dutch and Flemish patients, e.g., for

benchmarking purposes. Evaluating construct validity is vital to determine whether the bank is really measuring the intended construct. Absence of floor and ceiling effects is important for the discriminative and evaluative properties of an instrument.

The aim of the current study was to develop the DF-PROMIS-UE item bank v2.0, to investigate its cross-cultural and construct validity, as well as its floor and ceiling effects in Dutch patients with musculoskeletal UE disorders.

Methods

This study consisted of two parts: (1) the development of the DF-PROMIS-UE item bank v2.0 and (2) the evaluation of some of its psychometric properties. The development of the DF-PROMIS-UE item bank v2.0 consisted of a translation project that included cognitive debriefing interviews in order to check the comprehensibility and relevance of the preliminary item translations. The evaluation of some its measurement properties comprised evaluation of its cross-cultural and construct validity, and floor and ceiling effects.

Part 1: development

Translation

The translation of the PROMIS-UE items was integrated in a larger project to update the Dutch–Flemish PROMIS-PF (DF-PROMIS-PF) item bank from v1.2 (121 items) to v2.0 (165 items). All 45 newly developed PROMIS-PF items were translated into Dutch–Flemish, including the four new items of the DF-PROMIS-UE item bank v2.0. The translation process was performed similarly to the previous translation of Dutch–Flemish PROMIS item banks, using state of the art methodology [27–29]. In short, the process involved 2 forward translations (by 1 Dutch and 1 Flemish native-speaker), 1 reconciled version, 1 back translation by a native English speaker, comparison of original with back translation, and reviews by 3 bilingual experts (2 Dutch and 1 Flemish). Cognitive debriefing interviews were conducted for all 45 newly developed PROMIS-PF items.

Participants

Debriefing sample Consecutive eligible persons with ample knowledge of Dutch or Flemish were invited to participate in the cognitive debriefing interviews. A minimum of five native Dutch and five native Flemish patients, and five native Dutch and five native Flemish people from the general population, were invited to participate.

Part 2: evaluation measurement properties

Study design

A cross-sectional study design was used.

Participants

Dutch sample Patients who visited the outpatient clinic of the orthopedic department of the OLVG, a large teaching hospital (in Amsterdam, the Netherlands), were invited to participate in order to evaluate the measurement properties of DF-PROMIS-UE bank. Eligible patients were characterized as being 18 years or older with a musculoskeletal disorder of the UE, able to read and write in Dutch language, and to provide informed consent.

US sample Existing response data from persons from an US online panel, being 18 years or older, and having some difficulty due to UE pain or function, were also used to evaluate the cross-cultural validity of the Dutch–Flemish and US PROMIS-UE item banks [26]. More information about these persons is provided elsewhere [30].

Procedures

This part of the study was approved by the local institutional review boards of Slotervaart/Reade (Reference Number P1749) and the OLVG. Patients visiting the outpatient clinic of the orthopedic department between February and May 2018 were invited to fill in a web-based (digital) or paper-and-pencil (paper) questionnaire that included, among others, the DF-PROMIS-UE item bank.

Measures

First, the questionnaire included questions addressing demographic data, i.e., age, gender, country of birth, educational level, and clinical characteristics, i.e., location of pain, disease duration, and type of disorder.

Second, the questionnaire included the full DF-PROMIS-UE item bank v2.0. This bank measures the construct (domain) UE functioning, which is defined as activities that require use of the UE including the shoulder, arm and hand [31]. The bank contains 46 items. There are two different 5-point Likert scale response scales: (1) Unable to do/With much difficulty/With some difficulty/With a little difficulty/Without any difficulty; (2) Cannot do/Quite a lot/Somewhat/Very little/Not at all. No time-frame is specified, but current status is assumed. Higher scores indicate better function. The total score of the DF-PROMIS-UE item bank is expressed as a *T*-score, which is a standardized score, with 50 representing the average score of the US general population and 10 being its standard deviation (SD).

Third, the questionnaire included the Dutch–Flemish PROMIS Global Health Questionnaire v1.2. This questionnaire measures the overall evaluation of one's physical and mental health. It contains 10 items. There are two subscales; global physical health (GPH; 4 items) and global mental health (GMH; 4 items) [32]. The scores of the Dutch–Flemish PROMIS Global Health subscales are also expressed as *T*-scores. We used the Dutch–Flemish PROMIS pain intensity item (Global07r) from this bank as a legacy instrument for evaluating construct validity [32, 33]. It assesses pain intensity and consists of an 11-point numeric rating scale (NRS) with anchors 0 = "no pain" and 10 = "worst pain imaginable".

Fourth, the questionnaire contained three disease-specific legacy instruments:

1. The Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire, Subscale Disability/Symptoms, which measures physical function and symptoms in patients with musculoskeletal disorders of the upper limbs [3]. The subscale consists of 30 items. The time frame for the items is the past week. The total score ranges from 0 to 100, with higher scores indicating more disability. The DASH has satisfactory psychometric properties [4–6, 34, 35]. An official Dutch translation showed good psychometric properties [36, 37].
2. The Functional Index for Hand Osteoarthritis (FIHOA), that assesses functional impairment in patients with hand osteoarthritis. It consists of 10 items. No time frame is specified, but current status is assumed. Total scores range from 0 to 30, with higher scores indicating more functional impairment. The psychometric properties of the FIHOA are good [38–40]. An official Dutch translation showed good psychometric properties as well [41].
3. The Michigan Hand Outcomes Questionnaire (MHQ), subscale activities of daily living (MHQ-ADL), which assesses difficulty in performing daily activities for the right (5 items), the left (5 items) and both hands (7 items), in patients with conditions of, or injury to, the hand or wrist [42]. The time frame for the items is the past week. The MHQ-ADL total score is converted to a score from 0 to 100, with higher scores indicating less disability. The psychometric properties of the MHQ scale are good [42–51]. A Dutch translation of the MHQ showed good responsiveness [52].

Analysis

Demographic and clinical characteristics of the Dutch and US sample were summarized with descriptive statistics. Differences between the Dutch sample and the US sample were evaluated by χ^2-tests for categorical variables and independent sample-t-tests for continuous variables.

Cross-cultural validity of the DF-PROMIS-UE item bank was evaluated with differential item functioning (DIF) analyses. DIF analyses examine whether people from different groups (in this study: English and Dutch speaking patients) with the same level on the construct or trait (theta [θ], in this study: the UE function T-score) have different probabilities of giving a certain response to an item [16]. There are two types of DIF: uniform and non-uniform. Uniform DIF exists when the magnitude of DIF is constant across the trait. Non-uniform DIF exists when the magnitude of DIF varies across the trait, i.e., the item has a different discriminative ability in the groups. DIF for language was evaluated by ordinal logistic regression models with the item score as the dependent variable. An intercept model (Model 0) and three nested models were formed: Model 1 with theta as the explanatory variable, Model 2 with both theta and language as explanatory variables, and Model 3 with theta, language and an interaction term for language and theta as explanatory variables. A McFadden's pseudo R^2 change of 2% was used as the critical value to flag items with possible DIF [16, 53–55]. Items were flagged as having possibly non-uniform DIF, if the R^2 values of Models 2 and 3 differed by more than 2%, and possibly uniform DIF, if non-uniform DIF was absent and the R^2 values of Models 1 and 2 differed by more than 2%. If any items were flagged for DIF for language, the impact of DIF on the item scores was examined by plotting item characteristic curves (ICCs) and the impact on the DIF items on the test (total) score by plotting test characteristic curves (TCCs). The TCC plots show the test score for all 46 PROMIS-UE items and the test scores for the items flagged for DIF only [54, 55].

Construct validity was evaluated by calculating the correlations of the DF-PROMIS-UE item bank v2.0 T-scores with the total scores of the legacy instruments. Pearson's correlation coefficient r was used for normally distributed data and Spearman's correlation coefficients ρ for non-normally distributed data. Hypothesis were formulated a priori regarding the expected correlations according to the COSMIN guidelines [56, 57]. It was hypothesized that the DF-PROMIS-UE item bank would have a moderate negative correlation ($-0.50 < r \le -0.30$) with the Dutch–Flemish PROMIS pain intensity item [32, 33], given the fact that these instruments are intended to measure related constructs (UE physical function and pain, respectively) only. Moreover, we hypothesized that the DF-PROMIS-UE item bank would have strong negative correlations ($r \le -0.50$) with the DASH, Subscale Disability/Symptoms [3] and the FIHOA scores [38–40] and a strong positive correlation ($r \ge 0.50$) with the MHQ-ADL score [42], given the fact that these instruments are intended to measure the same construct (UE physical function).

To evaluate floor and ceiling effects, the proportions of patients who achieved the highest or lowest raw scores were calculated for each measure. These proportions were calculated for the full DF-PROMIS-UE item bank (raw scores 46 and 230, respectively) and the Short Form 7a (raw scores 7 and 35, respectively) in the 212 participants who completed all items. For all measures a floor effect referred to the proportion of patients with a poor health status whereas a ceiling effect referred to the proportion of patients with a good health status, and a proportion of 15% or more was considered a floor/ceiling effect [58]. We followed the international PROMIS standards with respect to the sample sizes for this study [59, 60]. These standards prescribe a minimum sample size of 200 participants for evaluating of DIF between language groups and a sample size of 50–100 participants for evaluating construct validity. DIF analyses were done with R using the package Lordif (version 0.3-3) whereas all other analyses were done with IBM SPSS Statistics 25 (Armork, New York, USA).

Results

Part 1: translation

Table 1 provides an overview of the translated PROMIS-UE items. A sufficient Dutch–Flemish translation was obtained for the four new items from the DF-PROMIS-UE item bank v2.0, and no separate translations for Dutch and Flemish were required.

In total, 28 native-speaking (18 Dutch and 10 Flemish) persons participated in cognitive debriefing interviews. Their mean age (standard deviation [SD]) was 46 (19) years, and 68% were female. Most participants were

Table 1 Description of the four new items of the Dutch–Flemish PROMIS upper extremity item bank v2.0

Item name	Description
PFM2	Are you able to lift a heavy painting or picture to hang on your wall above eye-level?
PFM16	Are you able to pass a 20-pound (10 kg) turkey or ham to other people at the table?
PFM18	Are you able to continuously swing a baseball bat or tennis racket back and forth for 5 minutes?
PFC8	Does your health now limit you in opening a previously opened jar?

patients with UE disorders (68%) whereas the remaining participants were healthy persons without complaints (32%).

During cognitive debriefing three out of four items (PFM2, PFM16 and PFM18) were considered to be less relevant or as describing unusual activities by some participants (both patients and people from the general population). Despite these comments, we decided to maintain the items without adaptation of the translation in the preliminary DF-PROMIS-UE item bank, enabling to investigate whether DIF for language would occur for these items.

Part 2: evaluation measurement properties

Participants

With respect to the Dutch sample 371 patients were screened for eligibility and 67 patients did not meet the selection criteria. Of the 304 patients fulfilling the selection criteria, 218 (72%) were willing to participate, provided informed consent, and completed the DF-PROMIS-UE item bank fully ($n = 212$) or partly ($n = 6$). Their data were used to study cross-cultural validity. Of the 304 patients fulfilling the selection criteria, 205 (67%) patients completed all measures, digitally ($n = 199$) or on paper ($n = 6$). Their data were used to study construct validity.

Table 2 summarizes the demographic and clinical characteristics of the Dutch and US samples. In the Dutch sample, the mean age was 53 years, half of them were female (50%), most were born in the Netherlands (73%) and had at least a high school degree (92%). Most patients reported having pain in one or both shoulder(s) (76%) or arm(s) (56%). Most reported to have a trauma (33%) or physical (e.g., muscle) injury (19%). The results of the t-test and χ^2 test showed that the Dutch participants, as compared to the US participants, were on average older, more often male, and differed in level of education.

Measures

Table 3 summarizes the scores on the DF-PROMIS-UE item bank, the PROMIS Global Health Questionnaire and the legacy instruments. The mean PROMIS-UE item bank T-scores of the Dutch sample (34.7 [SD = 8.6]) and the US sample (36.5 [SD = 7.0]) differed slightly, albeit statistically significant ($p < 0.05$, Hedges $g = 0.24$ [small]).

Cross-cultural validity

Table 4 summarizes the eight items that were flagged for DIF for language. Six items showed uniform DIF (PFA36,

PFB13, PFB21r1, PFB28r1, PFB56r1, and PFC43). Two items showed non-uniform DIF (PFM2 and PFM16) and the discrimination parameters were higher in Dutch patients than in US participants.

Figure 1 shows the impact of DIF for language in the TCC. The left graph shows the TCC for all 46 UE items, and the right graph shows the TCC for the eight items flagged as possibly having DIF only. The finding that the solid and the dashed curves in the left graph (all 46 UE items) are almost overlapping, indicates a minimal impact of DIF by language for the full item bank.

Construct validity

Table 5 summarizes the correlations between the DF-PROMIS-UE T-scores and the legacy instrument scores. All correlations were as hypothesized.

Floor and ceiling effects

Table 6 provides an overview of the proportion of participants that achieved the lowest or highest possible raw scores on the measures. No floor or ceiling effects were found for the DF-PROMIS-UE item bank, and no floor and a small (2.4%) ceiling effect for the Short Form 7a. No floor and a minimal (0.5%) ceiling effect were found for the DASH Subscale Disability/Symptoms, a minimal floor (0.5%) and a ceiling (17.6%) effect for the FIHOA, and no floor and some ceiling effect (11.7%) for the MHQ-ADL.

Discussion

The aim of this study was to develop the DF-PROMIS-UE item bank v2.0, to investigate its cross-cultural and construct validity, as well as its floor and ceiling effects in Dutch patients with musculoskeletal UE disorders. DIF analyses flagged eight items as possibly having DIF for language, but the impact of DIF on the test score was negligible, indicating sufficient cross-cultural validity. The construct validity for the item bank was sufficient, because none of the four predefined hypotheses about the correlations with legacy instruments had to be rejected. The full item bank and the short form had no floor or ceiling effects.

A limitation of our study is that the Dutch and US samples differed with respect to age, gender, educational level, administration mode, and the US sample was a non-clinical sample. These differences between the two samples might also have caused the DIF that we have found for the eight items. However, in previous studies, addressing the DF-PROMIS-PF item bank v1.2, that included 42 of the current 46 items, no DIF was found between groups differing with respect to age, gender, educational level and administration

Table 2 Demographic and clinical characteristics of the Dutch and US samples

Characteristic	Dutch sample (n=218)		US sample (n=246)		p-value
Age, mean (SD), range	53	(15), 18–87	48	(14), 18–85	<0.001[a]
Gender, n (%)					<0.001[b]
Male	109	(50)	76	(31)	
Female	109	(50)	170	(69)	
Country of birth, n (%)					
Netherlands	160	(73)	–		
Other	45	(21)	–		
Missing	13	(6)			
Social status, n (%)					
Single	69	(32)	–		
Married or living together	127	(58)	–		
Other	22	(10)	–		
Educational level, n (%)					<0.001[b]
Less than high school degree	18	(8)	6	(2)	
High school degree	21	(10)	53	(22)	
Some college	76	(35)	81	(33)	
College degree	14	(6)	80	(32)	
Advanced degree	89	(41)	26	(11)	
Employment status, n (%)					
Full time	84	(39)	–		
Part time	40	(18)	–		
Unpaid, volunteer	18	(8)	–		
Retired	40	(18)	–		
Unemployed	10	(5)	–		
Other	26	(12)	–		
Social benefits, n (%)					
Sick listed	40	(18)	–		
Disability benefit	21	(10)	–		
Unemployment benefit	5	(2)	–		
Other	18	(8)	–		
No social benefit	134	(62)	–		
Location of pain, n (%)[c]					
Shoulder(s)	165	(76)	–		
Arm(s)	122	(56)	–		
Hand(s)	47	(22)	–		
Fingers	42	(19)	–		
Other	41	(20)	–		
No pain	16	(7)	–		
Duration of complaints, n (%)					
<1 month	22	(10)	–		
1–3 months	22	(10)	–		
3–6 months	30	(14)	–		
6–12 months	36	(16)	–		
1–2 years	46	(22)	–		
2–5 years	31	(14)	–		
>5 years	31	(14)	–		
Type of disorder, n (%)[c]					
Physical injury (e.g., muscle)	41	(19)	–		
Overuse injury	32	(15)	–		
Trauma	72	(33)	–		
Surgery	20	(9)	–		

Table 2 (continued)

Characteristic	Dutch sample ($n=218$)		US sample ($n=246$)	p-value
Disease (e.g., arthritis)	6	(3)	–	
Congenital	5	(2)	–	
Other	39	(18)	–	
Unknown	39	(18)	–	

[a]Independent samples t-test

[b]χ^2 test

[c]Multiple answers allowed

Table 3 PROMIS and legacy instruments scores

	Dutch sample ($n=218$)			US sample ($n=246$)			p-value
	Mean	(SD)	Range	Mean	(SD)	Range	
PROMIS upper extremity item bank T-scores ($n=218$)	34.7	(8.6)	14.1–61.2	36.5	(7.0)	20.1–61.2	<0.05[a]
PROMIS Global Health Questionnaire T-scores ($n=205$)							
Subscale Global Physical Health	43.0	(7.1)	23.7–60.3	–			
Subscale Global Mental Health	47.0	(9.9)	24.1–67.6	–			
Legacy instruments total scores ($n=205$)							
PROMIS Global Health pain intensity[b]	4.9	(2.7)	0–10	–			
Disabilities of the Arm, Shoulder and Hand (DASH), Subscale Disability/Symptoms[c]	36.4	(20.9)	0–87.5	–			
Functional Index for Hand Osteoarthritis (FIHOA)[d]	7.0	(7.2)	0–30	–			
Michigan Hand Outcomes Questionnaire (MHQ), Subscale Activities of Daily Living[e]	74.3	(25.6)	0–100	–			

[a]Independent samples t-test

[b]Range 0–10, with higher scores indicating more pain

[c]Range 0–100, with higher scores indicating more disability

[d]Range 0–30, with higher scores indicating more functional impairment

[e]Range 0–100, with higher scores indicating less disability

Table 4 Dutch–Flemish PROMIS upper extremity items with differential item functioning (DIF) for language

Item name	English	Dutch–Flemish	McFadden's pseudo R^2 change comparing	
			Model 1–2	Model 2–3
Items with uniform differential item functioning				
PFA36	Are you able to put on and take of a coat or jacket?	Kunt u een jas of jasje aan- en uittrekken?	0.026	0.017
PFB13	Are you able to carry a shopping bag or briefcase?	Kunt u een boodschappentas of aktetas dragen?	0.026	0.001
PFB21r1	Are you able to pick up coins from a table top?	Kunt u munten van een tafel oppakken?	0.050	0.000
PFB28r1	Are you able to lift 10 lb (5 kg) above your shoulder?	Kunt u 5 kilo boven uw schouder tillen?	0.055	0.018
PFB56r1	Are you able to lift one pound (0.5 kg) to shoulder level without bending your elbow?	Kunt u 0,5 kilo tot op schouderhoogte tillen waarbij u uw elleboog gestrekt houdt?	0.033	0.002
PFC43	Are you able to use your hands, such as for turning faucets, using kitchen gadgets, or sewing?	Kunt u uw handen gebruiken, bijvoorbeeld om kranen open en dicht te draaien, keukengerei te gebruiken of te naaien?	0.040	0.000
Items with non-uniform differential item functioning				
PFM2	Are you able to lift a heavy painting or picture to hang on your wall above eye-level?	Kunt u een zwaar schilderij of fotolijst optillen om boven ooghoogte aan de muur te hangen?	0.017	0.024
PFM16	Are you able to pass a 20-pound (10 kg) turkey or ham to other people at the table?	Kunt u aan tafel een grote en zware schaal met eten (10 kilo) doorgeven aan een ander?	0.050	0.024

Fig. 1 The total impact of differential item functioning for language on the test characteristic curves (TCC). Both graphs show the relation between the level of upper extremity function (theta as estimated in the DIF analysis) on the *x*-axis and the total (raw) test scores on the *y*-axis. The left graph shows the TCC for all 46 Dutch–Flemish (DF [solid line]) and United States (US [dashed line]) PROMIS upper extremity items. The right graph shows the TCC for the eight items having DIF
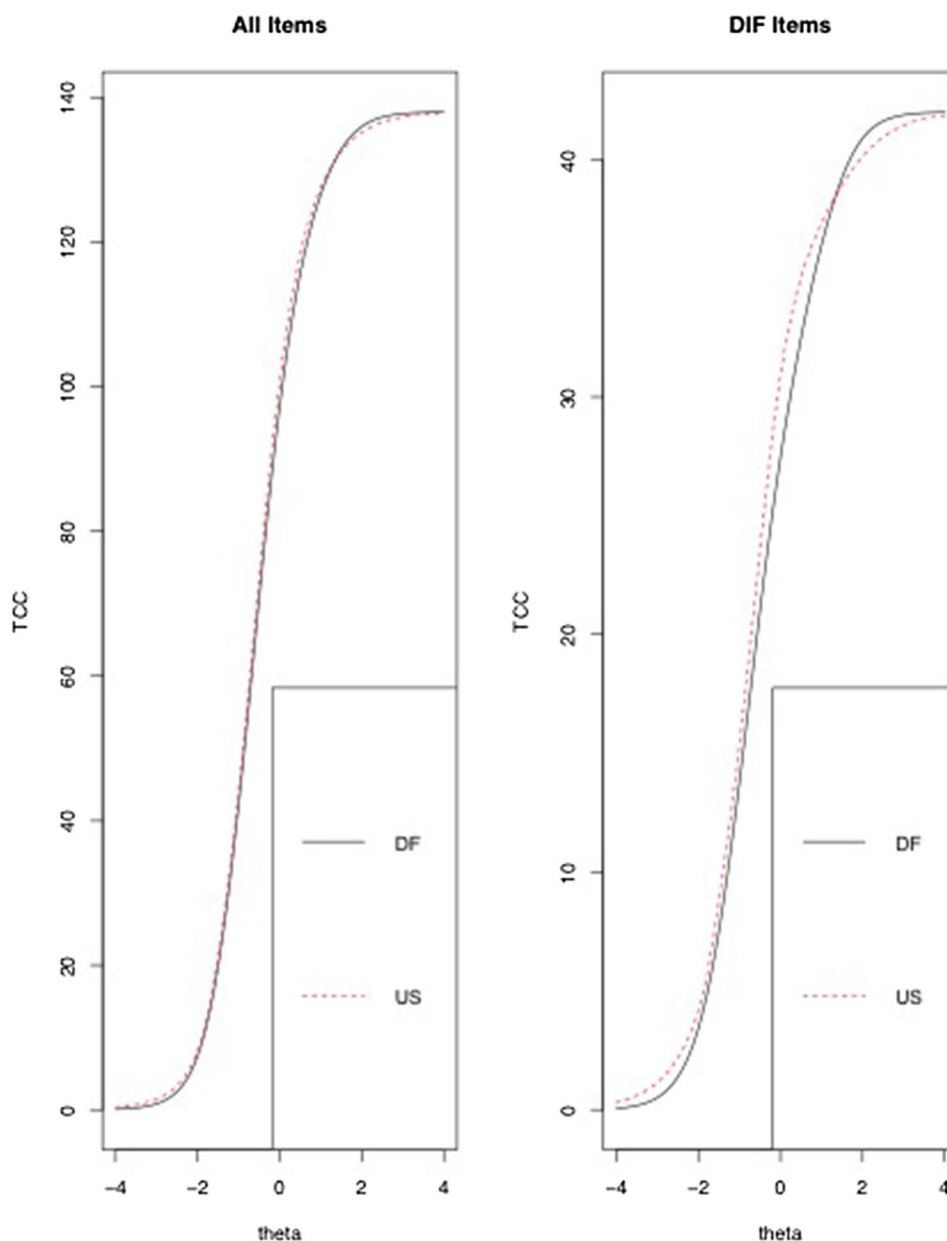

Table 5 Correlations between the Dutch–Flemish PROMIS upper extremity and legacy instruments

Legacy instrument	Hypothesized correlation	Observed correlation
Dutch–Flemish PROMIS pain intensity	$-0.50 < r \leq -0.30$	-0.43[a]
Disabilities of the Arm, Shoulder and Hand (DASH), Subscale Disability/Symptoms	$r \leq -0.50$	-0.87[b]
Functional Index for Hand Osteoarthritis (FIHOA)	$r \leq -0.50$	-0.86[b]
Michigan Hand Outcomes Questionnaire (MHQ), Subscale Activities of Daily Living	$r \geq 0.50$	0.87[b]

[a]Pearson's correlation coefficient r

[b]Spearman's correlation coefficients ρ

Table 6 Floor and ceiling effects of the full Dutch–Flemish PROMIS upper extremity item bank v2.0, its short form 7a and legacy instruments

	Floor, N (%)[a]	Ceiling, N (%)[b]	Completed, N
Dutch–Flemish PROMIS Bank v2.0: upper extremity[c]	0 (0)	0 (0)	212
Dutch–Flemish PROMIS Short Form v2.0: upper extremity 7a[d]	0 (0)	5 (2.4)	212
Disabilities of the Arm, Shoulder and Hand (DASH), Subscale Disability/Symptoms[e]	0 (0)	1 (0.5)	205
Functional Index for Hand Osteoarthritis (FIHOA)[f]	1 (0.5)	36 (17.6)	205
Michigan Hand Outcomes Questionnaire (MHQ), Subscale Activities of Daily Living[g]	0 (0)	24 (11.7)	205

[a] A floor effect refers to a poor health status

[b] A ceiling effect refers to a good health status

[c] Raw score range 46–230, with higher scores indicating better physical functioning. A score of 46 refers to a floor and a score of 230 to a ceiling effect

[d] Raw score range 7–35, with higher scores indicating better physical functioning. A score of 7 refers to a floor and a score of 35 to a ceiling effect

[e] Raw score range 0–100, with higher scores indicating more disability. A score of 100 refers to a floor and a score of 0 to a ceiling effect

[f] Raw score range 0–30, with higher scores indicating more functional impairment. A score of 30 refers to a floor and a score of 0 to a ceiling effect

[g] Raw score range 0–100, with higher scores indicating less disability. A score of 0 refers to a floor and a score of 100 to a ceiling effect

mode, and between several clinical samples and a non-clinical, general population, sample [61–64]. Therefore, it seems unlikely that the demographic and clinical differences, that we found between the Dutch and the American samples in this study, were an explanation for the DIF of the eight items. Nevertheless, we recommend, for future research, to study the DF-PROMIS-UE item bank v2.0 with respect to DIF for age, gender, educational level and administration mode, and between clinical and non-clinical samples. Moreover, the US sample used in this study was a subsample of the US calibration sample (and not the centering sample) [26]. If any bias exists between the US sample used in this study and the US centering sample, the results of our DIF analyses may be similarly biased.

This is the first study investigating the cross-cultural validity of the 46 item PROMIS-UE item bank v2.0 outside the US. Comparable to a study addressing the DF-PROMIS-PF item bank v1.2, we also found sufficient cross-cultural validity, although several items in both studies showed some DIF for language [61]. In the cognitive debriefing interviews, three items (PFM2, PFM16, and PFM18) were regarded as less relevant or as describing unusual activities by some participants. Two out of these three items, which reflect higher levels of UE function (PFM2 and PFM16) also showed non-uniform DIF and responses showed that the activities described in these items were more difficult for Dutch participants with lower levels of UE function. Four DIF items (PFA36, PFB13, PFB28r1, and again PFM16) are part of the standard 7a short form and PFM16 is the current starting item of the CAT algorithm. This might indicate that some items will be less suitable to maintain in the final DF-PROMIS-UE item bank or short form and that another starting item might be more appropriate for the Dutch–Flemish

CAT. This will have to be investigated in the final item bank calibration. Nevertheless, the right graph in Fig. 1 shows that, even if all eight items with DIF would be administered in a short form or CAT, the impact of DIF on the test score would be minimal. We therefore decided to keep these items in this preliminary version of the item bank.

To examine construct validity, we formulated a priori four hypotheses about the correlations with legacy instruments, as is proposed for studies on measurement instruments [56]. The constructs that are measured by the legacy instruments should be clear and these instruments should have sufficient measurement properties in a comparable population, which was the case in our study. None of the a priori formulated hypotheses were rejected, herewith indicating sufficient construct validity for the DF-PROMIS-UE item bank [57]. Three other studies examined the correlation between the US PROMIS-UE item bank v2.0 and legacy instruments. Minoughan and coworkers studied the bank, administered as a CAT, in patients with shoulder arthritis and found a moderate correlation ($r = 0.57$) with the American Shoulder and Elbow Surgeons (ASES) shoulder assessment form and a moderately strong correlation ($r = 0.64$) with the Simple Shoulder Test (SST) [65]. Kaat and colleagues reported, in a sample with participants with UE limitations, a correlation of 0.72 with the PROMIS-PF short form (SF8b), which is a generic physical function PROM, and a correlation of 0.69 with the Flexilevel Scale of Shoulder Function (FLEX-SF), which is a shoulder-specific PROM [26]. Van Bruggen and colleagues reported, in 303 patients from an outpatient department of a level 1 (academic) trauma center, correlations of the DF-PROMIS-UE item bank with the DASH, Patient-Reported Wrist Evaluation (PRWE) function and

MHQ-ADL of -0.84, -0.75, and -0.73, respectively. This study also showed a sufficient structural validity and internal consistency of the Dutch–Flemish PROMIS-UE item bank [66].

In previous studies, that examined the PROMIS-UE item bank (v1.2) in clinical populations with UE conditions, ceiling effects in the item bank were found [21–25]. In the current study, no floor or ceiling effect were found for the full DF-PROMIS-UE item bank, and no floor and a small, well below the 15% criterion, ceiling effect for the Short Form 7a. These findings are comparable to those in the study of Kaat et al. addressing the expansion and validation of the PROMIS-UE item bank v2.0 [26]. In our study, the FIHOA had a ceiling effect and the MHQ-ADL had some, below the 15% criterion, ceiling effect. These effects reduce the discriminatory and evaluating properties of a measure. Moreover, floor and ceiling effects may also exclude the application of some statistical analyses as many of them assume a normal distribution. Thus, the DF-PROMIS-UE item bank v2.0 has an improved measurement range compared to the initial PROMIS-UE item bank (v1.2) and the measurement range seems comparable to the US PROMIS-UE item bank v2.0.

In line with previous work of the Dutch–Flemish PROMIS group, the results of our study add to the evidence about the psychometric properties of the Dutch–Flemish PROMIS banks. Following the PROMIS guidelines, cross-cultural validation is the first recommended step after translation of PROMIS items banks [60]. Once cross-cultural validity has been established, further development of the item bank is warranted. We recommend to expand the current study to a larger sample, with a minimal sample size ≥ 500, for a so-called full item bank calibration. Afterwards, PROMIS CATs can be applied in clinical practice and research.

In conclusion, in this study we found sufficient cross-cultural and construct validity of the newly developed DF-PROMIS-UE item bank v2.0, and absence of floor and ceiling effects. Further validation of the item bank is now warranted and the item bank has the potential of improved measurement of UE functioning in the Dutch–Flemish population.

Compliance with ethical standards

References

1. Huisstede, B. M., Bierma-Zeinstra, S. M., Koes, B. W., & Verhaar, J. A. (2006). Incidence and prevalence of upper-extremity musculoskeletal disorders. A systematic appraisal of the literature. *BMC Musculoskeletal Disorders*. https://doi.org/10.1186/1471-2474-7-7.

2. Smith, E., Hoy, D. G., Cross, M., Vos, T., Naghavi, M., Buchbinder, R., et al. (2014). The global burden of other musculoskeletal disorders: Estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases, 73*(8), 1462–1469. https://doi.org/10.1136/annrheumdis-2013-204680.

3. Hudak, P. L., Amadio, P. C., & Bombardier, C. (1996). Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *American Journal of Industrial Medicine, 29*(6), 602–608.

4. Bot, S. D., Terwee, C. B., van der Windt, D. A., Bouter, L. M., Dekker, J., & de Vet, H. C. (2004). Clinimetric evaluation of shoulder disability questionnaires: A systematic review of the literature. *Annals of the Rheumatic Diseases, 63*(4), 335–341. https://doi.org/10.1136/ard.2003.007724.

5. Roy, J. S., MacDermid, J. C., & Woodhouse, L. J. (2009). Measuring shoulder function: A systematic review of four questionnaires. *Arthritis and Rheumatism, 61*(5), 623–632. https://doi.org/10.1002/art.24396.

6. Hoang-Kim, A., Pegreffi, F., Moroni, A., & Ladd, A. (2011). Measuring wrist and hand function: Common scales and checklists. *Injury, 42*(3), 253–258. https://doi.org/10.1016/j.injury.2010.11.050.

7. Forget, N. J., & Higgins, J. (2014). Comparison of generic patient-reported outcome measures used with upper extremity musculoskeletal disorders: Linking process using the International Classification of Functioning, Disability, and Health (ICF). *Journal of Rehabilitation Medicine, 46*(4), 327–334. https://doi.org/10.2340/16501977-1784.

8. Huang, H., Grant, J. A., Miller, B. S., Mirza, F. M., & Gagnier, J. J. (2015). A systematic review of the psychometric properties of patient-reported outcome instruments for use in patients with rotator cuff disease. *American Journal of Sports Medicine, 43*(10), 2572–2582. https://doi.org/10.1177/0363546514565096.

9. Thoomes-de Graaf, M., Scholten-Peeters, G. G., Schellingerhout, J. M., Bourne, A. M., Buchbinder, R., Koehorst, M., et al. (2016). Evaluation of measurement properties of self-administered PROMs aimed at patients with non-specific shoulder pain and "activity limitations": A systematic review. *Quality of Life Research, 25*(9), 2141–2160. https://doi.org/10.1007/s11136-016-1277-7.

10. Boyce, M. B., Browne, J. P., & Greenhalgh, J. (2014). 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 Quality and Safety, 23*(6), 508–518. https://doi.org/10.1136/bmjqs-2013-002524.

11. Rolstad, S., Adler, J., & Ryden, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value Health, 14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003.

12. Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F., et al. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research, 22*(8), 1889–1905. https://doi.org/10.1007/s11136-012-0344-y.

13. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–s11. https://doi.org/10.1097/01.mlr.0000258615.42478.55.

14. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011.

15. Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A., et al. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Quality of Life Research, 19*(9), 1311–1321. https://doi.org/10.1007/s11136-010-9694-5.

16. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5 Suppl 1), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.

17. Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*(Suppl 1), 133–141. https://doi.org/10.1007/s11136-007-9204-6.

18. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E., Jr. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526. https://doi.org/10.1016/j.jclinepi.2013.10.024.

19. Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology, 61*(1), 17–33. https://doi.org/10.1016/j.jclinepi.2006.06.025.

20. Hays, R. D., Spritzer, K. L., Amtmann, D., Lai, J. S., Dewitt, E. M., Rothrock, N., et al. (2013). Upper-extremity and mobility subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS) adult physical functioning item bank. *Archives of Physical Medicine and Rehabilitation, 94*(11), 2291–2296. https://doi.org/10.1016/j.apmr.2013.05.014.

21. Hung, M., Voss, M. W., Bounsanga, J., Crum, A. B., & Tyser, A. R. (2016). Examination of the PROMIS upper extremity item bank. *Journal of Hand Therapy*. https://doi.org/10.1016/j.jht.2016.10.008.

22. Beckmann, J. T., Hung, M., Voss, M. W., Crum, A. B., Bounsanga, J., & Tyser, A. R. (2016). Evaluation of the patient-reported outcomes measurement information system upper extremity computer adaptive test. *Journal of Hand Surgery American, 41*(7), 739–744.e734. https://doi.org/10.1016/j.jhsa.2016.04.025.

23. Anthony, C. A., Glass, N. A., Hancock, K., Bollier, M., Wolf, B. R., & Hettrich, C. M. (2017). Performance of PROMIS instruments in patients with shoulder instability. *American Journal of Sports Medicine, 45*(2), 449–453. https://doi.org/10.1177/0363546516668304.

24. Kaat, A. J., Rothrock, N. E., Vrahas, M. S., O'Toole, R. V., Buono, S. K., Zerhusen, T., Jr., et al. (2017). Longitudinal validation of the PROMIS physical function item bank in upper extremity trauma. *Journal of Orthopaedic Trauma, 31*(10), e321–e326. https://doi.org/10.1097/BOT.0000000000000924.

25. Beleckas, C. M., Padovano, A., Guattery, J., Chamberlain, A. M., Keener, J. D., & Calfee, R. P. (2017). Performance of Patient-Reported Outcomes Measurement Information System (PROMIS) Upper Extremity (UE) versus physical function (PF) computer adaptive tests (CATs) in upper extremity clinics. *Journal of Hand Surgery, 42*(11), 867–874. https://doi.org/10.1016/j.jhsa.2017.06.012.

26. Kaat, A. J., Buckenmaier, C. C., Cook, K. F., Rothrock, N. W., Schalet, B. D., Gershon, R. C., et al. (2019). The expansion and validation of a new upper extremity item bank for the Patient Reported Measurement Information System (PROMIS). *Journal of Patient-Reported Outcomes*. https://doi.org/10.1186/s41687-019-0158-6.

27. Terwee, C. B., Roorda, L. D., de Vet, H. C., Dekker, J., Westhovens, R., van Leeuwen, J., et al. (2014). Dutch–Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). *Quality of Life Research, 23*(6), 1733–1741. https://doi.org/10.1007/s11136-013-0611-6.

28. Eremenco, S. L., Cella, D., & Arnold, B. J. (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Evaluation and the Health Professions, 28*(2), 212–232. https://doi.org/10.1177/0163278705275342.

29. Bonomi, A. E., Cella, D. F., Hahn, E. A., Bjordal, K., Sperner-Unterweger, B., Gangeri, L., et al. (1996). Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. *Quality of Life Research, 5*(3), 309–320.

30. Gershon, R., & Kaat, A. (2019). PROMIS physical function upper extremity v2.0 extension (V1 ed.). Harvard Dataverse.

31. PROMIS. (2019). PROMIS physical function scoring manual. Retrieved May 27, 2019, from http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Physical_Function_Scoring_Manual.pdf.

32. Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research, 18*(7), 873–880. https://doi.org/10.1007/s11136-009-9496-9.

33. Sendlbeck, M., Araujo, E. G., Schett, G., & Englbrecht, M. (2015). Psychometric properties of three single-item pain scales in patients with rheumatoid arthritis seen during routine clinical care: A comparative perspective on construct validity, reproducibility and internal responsiveness. *RMD Open, 1*(1), e000140. https://doi.org/10.1136/rmdopen-2015-000140.

34. Changulani, M., Okonkwo, U., Keswani, T., & Kalairajah, Y. (2008). Outcome evaluation measures for wrist and hand: Which one to choose? *International Orthopaedics, 32*(1), 1–6. https://doi.org/10.1007/s00264-007-0368-z.

35. Schoneveld, K., Wittink, H., & Takken, T. (2009). Clinimetric evaluation of measurement tools used in hand therapy to assess activity and participation. *Journal of Hand Therapy, 22*(3), 221–235. https://doi.org/10.1016/j.jht.2008.11.005.

36. Veehof, M. M., Sleegers, E. J., van Veldhoven, N. H., Schuurman, A. H., & van Meeteren, N. L. (2002). Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). *Journal of Hand Therapy, 15*(4), 347–354.

37. De Smet, L., De Kesel, R., Degreef, I., & Debeer, P. (2007). Responsiveness of the Dutch version of the DASH as an outcome measure for carpal tunnel syndrome. *Journal of Hand Surgery, 32*(1), 74–76. https://doi.org/10.1016/j.jhsb.2006.10.001.

38. Klokker, L., Terwee, C. B., Waehrens, E. E., Henriksen, M., Nolte, S., Liegl, G., et al. (2016). Hand-related physical function in rheumatic hand conditions: A protocol for developing a

patient-reported outcome measurement instrument. *British Medical Journal Open, 6*(12), e011174. https://doi.org/10.1136/bmjopen-2016-011174.

39. Dreiser, R. L., Maheu, E., Guillou, G. B., Caspard, H., & Grouin, J. M. (1995). Validation of an algofunctional index for osteoarthritis of the hand. *Revue du Rhumatisme. English Edition, 62*(6 Suppl 1), 43S–53S.

40. Dreiser, R. L., Maheu, E., & Guillou, G. B. (2000). Sensitivity to change of the Functional Index for Hand Osteoarthritis. *Osteoarthritis Cartilage, 8*(Suppl A), S25–S28.

41. Wittoek, R., Cruyssen, B. V., Maheu, E., & Verbruggen, G. (2009). Cross-cultural adaptation of the Dutch version of the Functional Index for Hand Osteoarthritis (FIHOA) and a study on its construct validity. *Osteoarthritis Cartilage, 17*(5), 607–612. https://doi.org/10.1016/j.joca.2008.10.006.

42. Chung, K. C., Pillsbury, M. S., Walters, M. R., & Hayward, R. A. (1998). Reliability and validity testing of the Michigan Hand Outcomes Questionnaire. *Journal of Hand Surgery, 23*(4), 575–587. https://doi.org/10.1016/S0363-5023(98)80042-7.

43. Chung, K. C., Hamill, J. B., Walters, M. R., & Hayward, R. A. (1999). The Michigan Hand Outcomes Questionnaire (MHQ): Assessment of responsiveness to clinical change. *Annals of Plastic Surgery, 42*(6), 619–622.

44. Dias, J. J., Rajan, R. A., & Thompson, J. R. (2008). Which questionnaire is best? The reliability, validity and ease of use of the Patient Evaluation Measure, the Disabilities of the Arm, Shoulder and Hand and the Michigan Hand Outcome Measure. *Journal of Hand Surgery, 33*(1), 9–17. https://doi.org/10.1177/1753193407087121.

45. McMillan, C. R., & Binhammer, P. A. (2009). Which outcome measure is the best? Evaluating responsiveness of the Disabilities of the Arm, Shoulder, and Hand Questionnaire, the Michigan Hand Questionnaire and the Patient-Specific Functional Scale following hand and wrist surgery. *Hand (N Y), 4*(3), 311–318. https://doi.org/10.1007/s11552-009-9167-x.

46. Shauver, M. J., & Chung, K. C. (2009). The minimal clinically important difference of the Michigan hand outcomes questionnaire. *Journal of Hand Surgery, 34*(3), 509–514. https://doi.org/10.1016/j.jhsa.2008.11.001.

47. van de Ven-Stevens, L. A., Munneke, M., Terwee, C. B., Spauwen, P. H., & van der Linde, H. (2009). Clinimetric properties of instruments to assess activities in patients with hand injury: A systematic review of the literature. *Archives of Physical Medicine and Rehabilitation, 90*(1), 151–169. https://doi.org/10.1016/j.apmr.2008.06.024.

48. Chung, B. T., & Morris, S. F. (2014). Reliability and internal validity of the Michigan hand questionnaire. *Annals of Plastic Surgery, 73*(4), 385–389. https://doi.org/10.1097/SAP.0b013e31827fb3db.

49. London, D. A., Stepan, J. G., & Calfee, R. P. (2014). Determining the Michigan Hand Outcomes Questionnaire minimal clinically important difference by means of three methods. *Plastic and Reconstructive Surgery, 133*(3), 616–625. https://doi.org/10.1097/PRS.0000000000000034.

50. Chung, B. T., & Morris, S. F. (2015). Confirmatory factor analysis of the Michigan Hand Questionnaire. *Annals of Plastic Surgery, 74*(2), 176–181. https://doi.org/10.1097/SAP.0b013e3182956659.

51. Maia, M. V., de Moraes, V. Y., Dos Santos, J. B., Faloppa, F., & Belloti, J. C. (2016). Minimal important difference after hand surgery: A prospective assessment for DASH, MHQ, and SF-12. *SICOT Journal, 2*, 32. https://doi.org/10.1051/sicotj/2016027.

52. van der Giesen, F. J., Nelissen, R. G., Arendzen, J. H., de Jong, Z., Wolterbeek, R., & Vliet Vlieland, T. P. (2008). Responsiveness of the Michigan Hand Outcomes Questionnaire-Dutch language version in patients with rheumatoid arthritis. *Archives of Physical Medicine and Rehabilitation, 89*(6), 1121–1126. https://doi.org/10.1016/j.apmr.2007.10.033.

53. Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S115–S123. https://doi.org/10.1097/01.mlr.0000245183.28384.ed.

54. Choi, S., Gibbons, L. E., & Crane, P. K. (2018). Logistic ordinal regression differential item functioning using IRT, version 0.3-3. Retrieved May 30, 2018, from https://cran.r-project.org.

55. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo Simulations. *Journal of Statistical Software, 39*(8), 1–30.

56. Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., et al. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4.

57. Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., et al. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3.

58. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.

59. PROMIS. (2013). PROMIS instrument development and validation scientific standards, version 2.0 (revised May 2013). Retrieved May 27, 2019, from http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf.

60. PROMIS. (2014). Minimum requirements for the release of PROMIS instruments after translation and recommandations for futher psychometric evaluation. Retrieved May 27, 2019, from http://www.healthmeasures.net/images/PROMIS/Standards_for_release_of_PROMIS_instruments_after_translation_v8.pdf.

61. Crins, M. H. P., Terwee, C. B., Klausch, T., Smits, N., de Vet, H. C. W., Westhovens, R., et al. (2017). The Dutch–Flemish PROMIS physical function item bank exhibited strong psychometric properties in patients with chronic pain. *Journal of Clinical Epidemiology, 87*, 47–58. https://doi.org/10.1016/j.jclinepi.2017.03.011.

62. Oude Voshaar, M. A., ten Klooster, P. M., Glas, C. A., Vonkeman, H. E., Taal, E., Krishnan, E., et al. (2014). Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. *PLoS ONE, 9*(3), e92367. https://doi.org/10.1371/journal.pone.0092367.

63. Crins, M. H. P., van der Wees, P. J., Klausch, T., van Dulmen, S. A., Roorda, L. D., & Terwee, C. B. (2018). Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. *PLoS ONE, 13*(2), e0192187. https://doi.org/10.1371/journal.pone.0192187.

64. Crins, M. H. P., Terwee, C. B., Ogreden, O., Schuller, W., Dekker, P., Flens, G., et al. (2019). 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. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-2087-x.

65. Minoughan, C. E., Schumaier, A. P., Fritch, J. L., & Grawe, B. M. (2018). Correlation of PROMIS Physical Function Upper Extremity Computer Adaptive Test with American shoulder and elbow surgeons shoulder assessment form and simple shoulder test in patients with shoulder arthritis. *Journal of Shoulder and Elbow Surgery, 27*(4), 585–591. https://doi.org/10.1016/j.jse.2017.10.036.

66. van Bruggen, S. G. J., Lameijer, C. M., & Terwee, C. B. (2019). Structural validity and construct validity of the Dutch–Flemish PROMIS((R)) physical function-upper extremity version 2.0 item bank in Dutch patients with upper extremity injuries. *Disability and Rehabilitation*. https://doi.org/10.1080/09638288.2019.1651908.