


Psychometric Properties of the Pediatric Patient-Reported Outcomes Measurement Information System Item Banks in a Dutch Clinical Sample of Children With Juvenile Idiopathic Arthritis

Michiel A. J. Luijten,¹  Caroline B. Terwee,² Hedy A. van Oers,³ Mala M. H. Joosten,⁴ J. Merlijn van den Berg,³ Dieneke Schonenberg-Meinema,³ Koert M. Dolman,⁵ Rebecca ten Cate,⁶ Leo D. Roorda,⁷ Martha A. Grootenhuys,⁴ Marion A. J. van Rossum,⁸ and Lotte Haverman³

Objective. To assess the psychometric properties of 8 pediatric Patient-Reported Outcomes Measurement Information System (PROMIS) item banks in a clinical sample of children with juvenile idiopathic arthritis (JIA).

Methods. A total of 154 Dutch children (mean \pm SD age 14.4 \pm 3.0 years; range 8–18 years) with JIA completed 8 pediatric version 1.0 PROMIS item banks (anger, anxiety, depressive symptoms, fatigue, pain interference, peer relationships, physical function mobility, physical function upper extremity) twice and the Pediatric Quality of Life Inventory (PedsQL) and the Childhood Health Assessment Questionnaire (C-HAQ) once. Structural validity of the item banks was assessed by fitting a graded response model (GRM) and inspecting GRM fit (comparative fit index [CFI], Tucker-Lewis index [TLI], and root mean square error of approximation [RMSEA]) and item fit ($S-X^2$ statistic). Convergent validity (with PedsQL/C-HAQ subdomains) and discriminative validity (active/inactive disease) were assessed. Reliability of the item banks, short forms, and computerized adaptive testing (CAT) was expressed as the SE of theta (SE[θ]). Test-retest reliability was assessed using intraclass correlation coefficients (ICCs) and smallest detectable change.

Results. All item banks had sufficient overall GRM fit (CFI >0.95, TLI >0.95, RMSEA <0.08) and no item misfit (all $S-X^2 P > 0.001$). High correlations (>0.70) were found between most PROMIS T scores and hypothesized PedsQL/C-HAQ (sub)domains. Mobility, pain interference, and upper extremity item banks were able to discriminate between patients with active and inactive disease. Regarding reliability, PROMIS item banks outperformed legacy instruments. Post hoc CAT simulations outperformed short forms. Test-retest reliability was strong (ICC >0.70) for all full-length item banks and short forms, except for the peer relationships item bank.

Conclusion. The pediatric PROMIS item banks displayed sufficient psychometric properties for Dutch children with JIA. PROMIS item banks are ready for use in clinical research and practice for children with JIA.

INTRODUCTION

In recent years, the focus of health care has been drifting toward the inclusion of health-related quality of life (HRQoL) outcomes for patients in research and daily clinical practice by

administering patient-reported outcome measures (PROMs) (1–6). Previous studies have shown that rheumatology could benefit greatly from the use of patient-reported outcomes, as patients experience a wide array of problems (7) for which there is a disconnect between patient-reported outcomes and outcomes reported

Supported by Pfizer Pharmaceuticals (grant WP 465659).

¹Michiel A. J. Luijten, MSc: Emma Children's Hospital, Amsterdam University Medical Center, University of Amsterdam, Vrije Universiteit Amsterdam, and Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; ²Caroline B. Terwee, PhD: Amsterdam University Medical Center, Vrije Universiteit Amsterdam, and Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; ³Hedy A. van Oers, MSc, J. Merlijn van den Berg, MD, PhD, Dieneke Schonenberg-Meinema, MD, Lotte Haverman, PhD: Emma Children's Hospital, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands; ⁴Mala M. H. Joosten, MSc, Martha A. Grootenhuys, PhD: Princess Maxima Centre for Pediatric Oncology, Utrecht, The Netherlands; ⁵Koert M. Dolman, MD, PhD: Amsterdam Rheumatology and Immunology Centre, Reade, and Onze Lieve Vrouwe Gasthuis West, Amsterdam,

The Netherlands; ⁶Rebecca ten Cate, MD, PhD: Leiden University Medical Centre, Leiden, The Netherlands; ⁷Leo D. Roorda, MD, PT, PhD: Amsterdam Rehabilitation Research Center, Reade, Amsterdam, The Netherlands; ⁸Marion A. J. van Rossum, MD, PhD: Amsterdam Rheumatology and Immunology Centre, Reade, and Emma Children's Hospital, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

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

Address correspondence to Lotte Haverman, PhD, Psychosocial Department, Emma Children's Hospital, Amsterdam University Medical Center, Meibergdreef 9, Postbus 22660, 1100 AD Amsterdam, The Netherlands. Email: l.haverman@amc.nl.

Submitted for publication May 22, 2019; accepted in revised form October 15, 2019.

SIGNIFICANCE & INNOVATIONS

- This article provides an extensive overview of the psychometric properties of the Dutch pediatric Patient-Reported Outcomes Measurement Information System (PROMIS) item banks in a sample of children with juvenile idiopathic arthritis (JIA).
- This is the first study to provide a full calibration of Dutch PROMIS pediatric item banks in a clinical sample.
- This article demonstrates the advantages of computerized adaptive testing in clinical populations such as children with JIA.

by parents or clinicians (8). In clinical practice, there are often multiple PROMs available to measure the same construct/domain that differ in content, length, and scoring methods. These PROMs vary in their psychometric quality and often suffer from ceiling or floor effects when assessing patients who are outside the measurement range of the questionnaire. Most traditional PROMs (also known as legacy instruments) are scored using classical test theory (CTT), where all questions carry the same weight when calculating domain scores. The domain scores of these PROMs are incomparable due to the ordinal scoring methods used in CTT. In item response theory (IRT) modeling, the difficulty and discriminatory power of items can be taken into account when calculating a domain score. Additionally, IRT uses interval-based scores, which allows comparison of scores on the same metric. Therefore, a group of researchers from several US-based academic institutions and the National Institutes of Health initiated the creation of the Patient-Reported Outcomes Measurement Information System (PROMIS) (9,10), a new, universal set of IRT-based PROMs for adults and children that can accurately and quickly assess aspects of physical, mental, and social health of patients (9,11).

The US PROMIS group developed several item banks to assess relevant domains of physical, mental, and social health, such as fatigue, pain interference, or peer relationships (10). An item bank is a collection of a large number of items intended to measure 1 construct over a wide range of functioning, symptoms, or evaluations of well-being. This allows comparisons between different samples using the same PROM. The PROMIS item banks were developed using IRT modeling, which allows us to order items based on their difficulty. Using this information, items can be selected from the full-length item bank to create a short form, which measures a similar range of functioning as the full-length item bank. An online alternative to short forms is computerized adaptive testing (CAT). CAT uses the information of the IRT model (i.e., item difficulty and discrimination) and previous responses (11) to choose which items to administer to a specific patient. If, for example, a patient answers that he or she is never tired, the CAT will not offer an item about being exhausted to this patient, as the item about being exhausted has a higher difficulty. CAT thus provides more tailored items to patients than short forms, which

makes the estimates of the construct more reliable (12). As long as items are selected from the same item bank, scores from short forms and CATs can be compared on the same scale.

In 2009, the Dutch-Flemish PROMIS group (www.dutchflemishpromis.nl) was founded, followed by the Dutch-Flemish pediatric PROMIS group in 2011, to translate and implement the PROMIS item banks in The Netherlands and Flanders, Belgium. The pediatric PROMIS group translated 9 full PROMIS item banks into Dutch-Flemish (13).

The goal of this study was to assess the psychometric properties of 8 Dutch-Flemish PROMIS pediatric item banks in a clinical sample of Dutch children with juvenile idiopathic arthritis (JIA). The application of PROMIS is highly anticipated within rheumatology (14,15), and psychometric properties of the pediatric item banks were previously assessed in children with JIA in the US (8), making comparisons possible.

In the current study, the structural validity of the item banks was investigated and construct validity was assessed by comparing the PROMIS instruments to legacy instruments (the Pediatric Quality of Life Inventory [PedsQL] and the Childhood Health Assessment Questionnaire [C-HAQ]) and by comparing scores from patients with active and inactive disease. Furthermore, we assessed the reliability of the individual measurements for full-length item banks, short forms, and CATs. Finally, we assessed the test-retest reliability of the PROMIS item banks.

PATIENTS AND METHODS

Participants. All children diagnosed with JIA, 8–18 years of age, and under treatment in the Emma Children's Hospital Amsterdam University Medical Centers, Onze Lieve Vrouwe Gasthuis West, the Reade center for Rehabilitation and Rheumatology in Amsterdam, or the Leiden University Medical Centre in Leiden, were eligible and asked to participate in the study between June 2015 and January 2017. The study was approved by the medical ethics committees of all the participating centers. An invitation was sent to children and their parents to log in to the study website (www.hetkikt.nu/promis). All participants provided informed consent. Participating children were asked to complete 8 full pediatric PROMIS item banks at the start of the study (T1) and again 10 days later (T2) to assess test-retest reliability. Additionally, participants were asked to complete the PedsQL and C-HAQ at T1. All questionnaires were completed online. A reminder for T1 and T2 was sent out 3 days after the initial invitation. Children unable to understand Dutch or children with limitations/disorders that made them unable to complete (online) questionnaires were excluded from the study. Nonrespondent data were not available.

Patient characteristics. Personal data on age and sex were provided by the children. Medical data on the type of JIA, presence of uveitis, medication use, age at disease onset, disease duration, physician score of disease activity, and the number of joints with arthritis (1 = monoarthritis, 2–4 = oligoarthritis,

5–10 = polyarthritis, >10 = severe polyarthritis) were extracted retrospectively by a pediatric rheumatology expert (MAJvR) from the electronic medical records. The type of JIA was categorized in accordance to the International League of Associations for Rheumatology criteria (16). Disease activity was extracted from medical records by a rheumatologist (MAJvR) using the 100-mm physician visual analog scale (VAS; range 0–100, with 0 indicating no disease activity and higher scores indicating more activity).

Measures. *PROMIS item banks.* Eight full-length, Dutch PROMIS, version 1.0, pediatric self-report item banks (anger [17], anxiety [18], depressive symptoms [18], fatigue [19], pain interference [20], peer relationships [21], physical function mobility, and physical function upper extremity [22]) were completed by the children. All item banks utilize a 7-day recall period. A 5-point Likert scale ranging from 1 (“never”) to 5 (“almost always”) is used for all item banks, except the mobility and upper extremity item banks. For these item banks, the response categories range from 1 (“not able to do”) to 5 (“with no trouble”). Total scores are calculated by applying the original US IRT model to the data and estimating the level of functioning of the patient (theta). This level of functioning is transformed into a T score, with a score of 50 representing the mean of the general US population (SD 10). For all item banks, higher scores represent more of the construct (e.g., better mobility or more pain interference). Scores can also be calculated for the standard PROMIS short forms, consisting of 8 items for all domains, except for anger (5 items) and fatigue (10 items), by extracting short-form item responses from the full-length item bank.

PedsQL generic scale 4.0. The PedsQL (23) is a 23-item questionnaire that assesses the self-reported HRQoL of children (ages 8–18 years) across the following 4 domains: physical functioning (8 items); emotional functioning (5 items); social functioning (5 items); and school functioning (5 items). The PedsQL utilizes a 7-day recall period. Items are scored using a 5-point Likert scale ranging from 1 (“never a problem”) to 5 (“almost always a problem”). The response options are transformed into values of 0, 25, 50, 75, and 100, respectively. Domain scores (range 0–100, with a higher score representing better functioning) are calculated by summing and averaging the items within each domain. The total PedsQL score (range 0–100) is calculated by averaging all individual item scores. The PedsQL is an often used, validated tool for Dutch children with JIA (7,24).

C-HAQ. The C-HAQ is a 30-item questionnaire that measures self-reported functional ability in children (ages 8–18 years) (25). The C-HAQ is composed of the following 8 categories: dressing and grooming (4 items); arising (2 items); eating (3 items); walking (2 items); hygiene (5 items); reach (4 items); grip (5 items); and activities (5 items). The C-HAQ utilizes a 1-week recall period. Each item on the C-HAQ is scored from 0 (“without any difficulty”) to 3 (“unable to do”). The highest scoring item within a category determines the score for that category.

The disability index (range 0 [low]–3 [high]) averages the category scores. Additionally, the C-HAQ contains two 100-mm VAS to measure pain (0 = no pain, 100 = very severe pain) and well-being (0 = very well, 100 = very poor) over the past week. The C-HAQ is a validated tool for assessing Dutch children with JIA (25,26) and a recommended instrument for assessing daily functioning in rheumatology patients (27).

Statistical analysis. *Patient characteristics.* Descriptive analyses were performed to describe sociodemographic and clinical characteristics of the children, using SPSS, version 24.0 (28). All further analyses were performed in R (29).

Structural validity. To assess the structural validity of the PROMIS item banks, a graded response model (GRM) was fitted to each of the item banks. A GRM is an IRT model for items with ordinal response categories and requires several assumptions to be met, such as unidimensionality, local independence, and monotonicity. To assess unidimensionality of each item bank, a confirmatory factor analysis (CFA) was performed using the R-package lavaan, version 0.6-3 (30). An acceptable fit of a unidimensional model is indicated by a comparative fit index (CFI) value and Tucker-Lewis Index (TLI) score >0.95, a standardized root mean square residual (SRMR) value <0.10, and a root mean square error of approximation (RMSEA) value <0.08 (31). Scaled indices were reported. Local independence was assessed by looking at the residual correlations in the CFA model. An item pair is considered local independent when it has a residual correlation <0.20 (32). Finally, monotonicity was assessed using Mokken scaling (33,34). The assumption of monotonicity is met when the item H values of all items are ≥ 0.30 and the H value of the entire scale is ≥ 0.50 .

Once the assumptions were met, a GRM was fitted to each item bank to estimate item discrimination and threshold parameters using the expectation-maximization algorithm within the R-package mirt, version 1.29 (35). The discrimination parameter (α) represents the ability of an item to distinguish between patients with a different level of functioning (θ). The threshold parameters (β) represent the required level of functioning of a person to choose a higher response category over a lower response category. Previous simulation studies have shown that fitting a GRM requires a large sample size of ~500 respondents in most cases, but that increased unidimensionality and high discriminatory parameters of an item bank reduce the number of respondents required (36,37). As the items in PROMIS item banks were specifically chosen based on their discriminatory power and their contribution to measuring a single construct, we expected that a smaller sample size could be used. Caution is advised when assessing the estimated parameters, however, as other sample characteristics (i.e., skewness of responses) can impact parameter calibration. Model fit of the GRM model was assessed using the same CFI, TLI, SRMR, and RMSEA criteria as for the CFA. Item fit was assessed using the $S-X^2$ statistic (38), which calculates the differences between observed

and expected responses under the GRM model. A P value of the $S-X^2$ statistic <0.001 for an item is considered an item misfit (32).

Construct validity. Construct validity was investigated by assessing convergent and discriminative validity. Convergent validity was assessed by correlating the PROMIS item bank T scores to the PedsQL or C-HAQ using Pearson's correlation coefficient (r). A strong correlation (>0.70 or lower than -0.70) was expected between PROMIS T scores and the sum scores of the PedsQL and C-HAQ scales measuring similar constructs. Correlations with unrelated constructs were expected to be lower ($\Delta r > 0.10$).

Discriminative (known-groups) validity was assessed by comparing the T scores of PROMIS item banks between patients with an active and inactive disease using an independent sample t -test. Disease activity can be represented by results from the physician VAS and the number of joints with arthritis. However, a combination of these variables would result in an active disease group too small for valid comparison. The correlation between these 2 variables was high ($r = 0.75$), indicating that a combination of these variables would not impact the results much. Therefore, the physician VAS was used to discriminate active (>0) and inactive (0) disease, as this resulted in large enough groups for valid and reliable comparisons. It was expected that the physical health domains would be most affected by JIA (7,24). Mobility and upper extremity T scores were hypothesized to be significantly lower for patients with an active disease. The pain interference T scores were expected to be significantly higher for patients with active disease. For the remaining item banks, no differences in T scores were hypothesized between patients with active and inactive disease. Each PROMIS item bank was considered to have sufficient construct validity if at least 75% of the hypotheses were confirmed.

Reliability. In IRT, the reliability of an item bank can vary across levels of the measured construct. The estimated level of functioning is represented by θ , which is standardized to have a mean of 0 and an SD of 1 in the calibration sample. Each response pattern has a θ estimate and an associated SE of theta estimate ($SE(\theta)$). An $SE(\theta)$ of 0.32 corresponds to a reliability of 0.90. To compare the reliability of the PROMIS item banks to similar domains on the PedsQL, a GRM was fit to each PedsQL domain to calculate the θ estimates and $SE(\theta)$. Thetas and $SE(\theta)$ s were estimated for the full-length PROMIS item banks and short forms using the expected a posteriori (EAP) estimator. Post hoc CAT simulations were performed using R-package *catR*, version 3.16 (39) for each item bank, using maximum posterior weighted information selection criterion and the EAP estimator (40) to assess whether or not a CAT would outperform short forms. The starting item for each CAT was the item that offered most information at the mean of the population ($\theta = 0$). The maximum number of items for the CAT simulation was set to the number of items in the short form of the same item bank, which ensured that the CAT did not administer more items than the short form. The stopping rule of the $SE(\theta)$ was <0.32 (41).

Table 1. Patient characteristics*

Characteristics	No.	Value
Age, mean \pm SD years	157	14.4 \pm 3.0
Age at onset of JIA, mean \pm SD years	157	8.9 \pm 4.5
Sex, female	111	70.7
JIA subtype		
Oligoarticular JIA, persistent	26	16.6
Oligoarticular JIA, extended	16	10.2
Polyarticular JIA, RF negative	62	39.5
Polyarticular JIA, RF positive	7	4.5
Enthesitis-related arthritis	21	13.4
Psoriatic arthritis	11	7.0
Undifferentiated arthritis	0	0
Systemic JIA	4	2.5
Chronic arthritis with other autoimmune inflammatory disease	8	5.1
Disease specifications		
Disease duration, median (range)	157	4.9 (0.18–16.8)
Physician assessment of disease activity, VAS score (range 0–100)†	140	0 (0–50)
Number of joints with arthritis‡		
No arthritis	119	75.8
Monoarthritis (1 joint)	14	8.9
Oligoarthritis (2–4 joints)	11	7.0
Polyarthritis (>4 joints)	5	3.2
Presence of uveitis	26	16.6
Medication at time point of evaluation		
No medication	60	39.0
NSAIDs	20	12.7
MTX	69	43.9
Other DMARDs	4	2.5
Anti-TNF	45	28.7
Other biologics	2	1.3
Multiple medications	38	24.2

* Values are the percentage unless indicated otherwise. JIA = juvenile idiopathic arthritis; RF = rheumatoid factor; VAS = visual analog scale; NSAIDs = nonsteroidal antiinflammatory drugs; MTX = methotrexate; DMARDs = disease-modifying antirheumatic drugs; anti-TNF = anti-tumor necrosis factor.

† Physician VAS outcomes were missing for 17 patients at the time of measurement.

‡ Information on the number of infected joints was missing for 8 patients at the time of measurement.

Test-retest reliability. Test-retest reliability was assessed for the full-length item banks and the short forms by calculating the intraclass correlation coefficient (ICC; two-way random-effects model for absolute agreement) (42) of the T scores for the patients who completed the PROMIS item banks twice (within 4 weeks). An ICC >0.70 was considered acceptable (42). The smallest detectable change (SDC) was calculated for all full-length item banks as $1.96 \times \sqrt{2} \times (SD \times [\sqrt{1 - ICC}])$. The SDC represents the smallest change in score that falls outside of the measurement error (42).

RESULTS

Patient characteristics. A total of 154 children with JIA completed all PROMIS pediatric item banks, the PedsQL, and the C-HAQ. A total of 111 children completed the item banks twice, with a time interval ranging 1–14 weeks (mean 2.6). Patient characteristics are shown in Table 1. The mean \pm SD

Table 2. Model assumptions of the PROMIS pediatric item banks for children with juvenile idiopathic arthritis (n = 155)*

Item bank	Unidimensionality				Local independence, no. (%)†	Monotonicity, H scale
	CFI score	TLI score	SRMR	RMSEA		
Anger scale	0.995	0.989	0.032	0.053	0 (0)	0.726
Anxiety	0.983	0.980	0.077	0.103	1 (1.3)	0.662
Depressive symptoms	0.996	0.995	0.035	0.000	0 (0)	0.733
Fatigue	0.991	0.990	0.042	0.055	0 (0)	0.743
Mobility (n = 156)	0.992	0.991	0.072	0.000	6 (2.4)	0.588
Pain interference	0.987	0.985	0.044	0.059	0 (0)	0.682
Peer relationships	0.954	0.947	0.080	0.080	4 (3.8)	0.508
Upper extremity (n = 157)	0.991	0.990	0.073	0.021	5 (1.2)	0.573

* PROMIS = Patient-Reported Outcomes Measurement Information System; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation.

† Locally dependent item pairs.

age of patients was 14.4 ± 3.0 years (range 8–18 years), and the majority of the patients were female (70.7%). The majority of patients were diagnosed with polyarticular JIA (44.0%). More than one-half of the patients had inactive disease (66.9%), as measured by the physician VAS (n = 140). The distribution of joints affected by arthritis (n = 149) was 75.8% no arthritis, 8.9% monoarthritis, 7.0% oligoarthritis, and 3.2% polyarthritis.

Structural validity. Unidimensionality was sufficient for all item banks except for the anxiety item bank (RMSEA = 0.103) (Table 2). Local independence did not hold for all item banks (not for anxiety, mobility, peer relationships, and upper extremity). As the percentages of local dependent item pairs were low (1–4%), the GRM analyses were performed without removing items. The assumption of monotonicity was met for all items and item banks. The item parameters and item fit statistics of the fitted GRMs are available in Supplementary Table 1, available on the *Arthritis Care & Research* website at <http://onlinelibrary.wiley.com/doi/10.1002/acr.24094/abstract>. The discrimination parameters

ranged from 1.07 to 22.25. Two discrimination parameters of the upper extremity item bank had outlying discriminatory values ($\alpha > 10$). For the item banks peer relationships, mobility, and upper extremity, not all item thresholds could be calculated, as not all response categories were used by the respondents. There were no items with item misfit ($S-X^2 < 0.001$) in any of the item banks.

Construct validity. The correlations between the PROMIS item banks, the PedsQL, and the C-HAQ are shown in Table 3. For all item banks, at least 1 expected strong correlation (>0.70) with a relevant PedsQL or C-HAQ subdomain was found, except for the peer relationship item bank. For the mobility and upper extremity item banks, additional correlations were found that were nearly the same strength ($\Delta r < 0.10$) as the hypothesized strong correlation with the PedsQL physical subscale.

Discriminative validity was assessed by comparing T scores from patients with active disease (n = 35) to those from patients with inactive disease (n = 105). The results are shown

Table 3. Convergent and discriminative validity of the pediatric PROMIS item banks for children with juvenile idiopathic arthritis (n = 154)*

PROMIS questionnaire item	Convergent validity, PedsQL			Convergent validity, C-HAQ			Discriminant validity			Total hypotheses correct, %
	Physical	Emotional	Social	Total	Grip	Pain	Active disease†	Inactive disease‡	Mean difference \pm SD	
Anger	-0.48	-0.72§	-0.58	0.48	0.46	0.37	50.60	49.63	-0.97 \pm 1.86	100
Anxiety	-0.50	-0.78§	-0.62	0.48	0.46	0.38	49.64	50.28	0.64 \pm 1.81	100
Depressive symptoms	-0.54	-0.79§	-0.60	0.48	0.42	0.48	51.25	49.51	-1.74 \pm 1.84	100
Fatigue	0.76§	-0.62	-0.61	0.61	0.49	0.67	51.86	49.22	-2.64 \pm 1.91	86
Mobility	0.83§	0.52	0.67	-0.74§	-0.52	-0.71§	46.58	51.2	4.62 \pm 1.85¶	71
Pain interference	-0.76§	-0.62	-0.65	0.64§	0.49	0.75§	53.36	48.43	-4.93 \pm 1.83¶	86
Peer relationships	0.29	0.44	0.54§	-0.32	-0.28	-0.22	51.17	49.72	-1.45 \pm 1.90	71
Upper extremity	0.79§	0.56	0.65	-0.77§	-0.70§	-0.66§	47.37	51.18	3.80 \pm 1.75¶	71

* PROMIS = Patient-Reported Outcomes Measurement Information System; PedsQL = Pediatric Quality of Life Inventory; C-HAQ = Childhood Health Assessment Questionnaire.

† N = 35.

‡ N = 103.

§ Significant; numbers were hypothesized to be highly (>0.70) correlated or able to discriminate between patients with active and inactive disease.

¶ Significant at $P < 0.05$; numbers were hypothesized to be highly (>0.70) correlated or able to discriminate between patients with active and inactive disease.

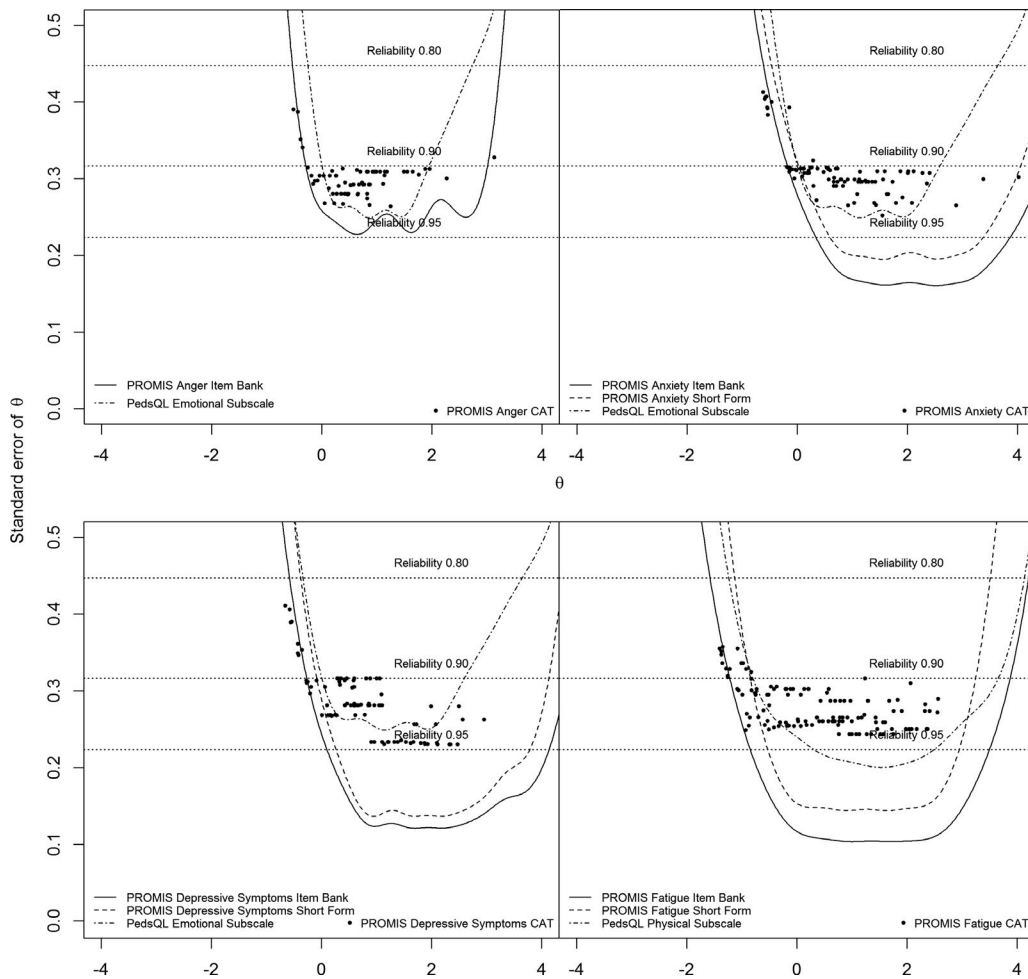


Figure 1. Reliability of measurements (expressed as SE of theta) of the full item bank/scale, short forms, post hoc computerized adaptive testing (CAT), and their related subdomain from the Pediatric Quality of Life Inventory (PedsQL) across the range of theta for the domain (top left to bottom right): Anger, Anxiety, Depressive Symptoms, and Fatigue. PROMIS = Patient-Reported Outcomes Measurement Information System.

in Table 3. Patients with active disease scored significantly lower on the mobility item bank (mean difference -4.62 ; $t(138) = 2.50$, $P = 0.014$) and the upper extremity item bank (mean difference -3.81 ; $t(137) = 2.17$, $P = 0.032$) than patients with inactive disease. For the pain interference item bank, patients with active disease scored significantly higher (mean difference 4.93 ; $t(136) = -2.70$, $P = 0.008$) than patients with no disease activity.

For the anger, anxiety, depressive symptoms, fatigue, and pain interference item banks, at least 75% of the hypotheses regarding construct validity were confirmed. The mobility, upper extremity, and peer relationships item banks did not meet the criterion (71%).

Reliability. All PROMIS item banks provided reliable measurements ($SE[\theta] < 0.32$) for the sample mean of 0 and a range of at least 2 SD of theta in the direction of clinical interest (e.g., higher thetas for depressive symptoms, lower thetas for mobility). The only exception was the upper extremity item bank, which did not reach satisfactory reliability for the mean. The reliability of measurements

of the full item bank, short forms, post hoc CATs, and their related subdomain from the PedsQL across the range of theta for all items banks is visualized in Figures 1 and 2. The number of reliable measurements, the number of items used, and the average $SE(\theta)$ value of the full item banks, short forms, and CATs are shown in Table 4.

Test-retest reliability. Ten patients were removed from the test-retest reliability analyses, as they did not complete the second measurement within 4 weeks of the initial measurement. Most item banks displayed sufficient ($ICC > 0.70$) test-retest reliability. Only the item bank peer relationships displayed a moderate test-retest reliability ($ICC 0.69$). The SDC ranged from 12.1 to 18.7. The SDC and ICC values are shown in Table 4.

DISCUSSION

This is the first study to assess the psychometric properties of the pediatric PROMIS item banks in a Dutch clinical sample.

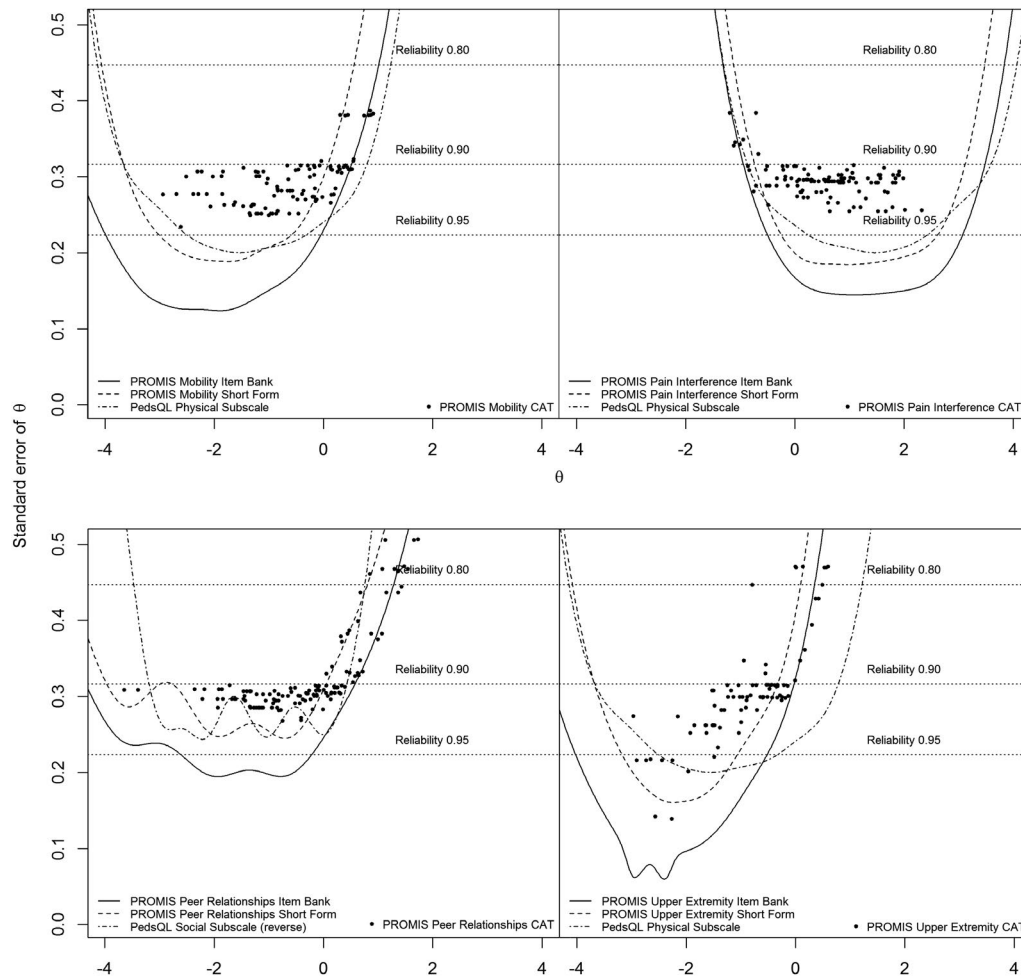


Figure 2. Reliability of measurements (expressed as SE of theta) of the full item bank/scale, short forms, post hoc computerized adaptive testing (CAT), and their related subdomain from the Pediatric Quality of Life Inventory (PedsQL) across the range of theta for the domain (top left to bottom right): Mobility, Pain Interference, Peer Relationships, and Upper Extremity. PROMIS = Patient-Reported Outcomes Measurement Information System.

The PROMIS item banks all displayed sufficient validity and reliability for use in clinical practice for children with JIA. All item banks fit the underlying IRT model. The item banks correlated highly with similar (sub)domains from the legacy instruments PedsQL and C-HAQ. The item banks pain interference, mobility, and upper extremity were able to discriminate between active and inactive JIA. Other studies have shown that issues regarding physical health commonly occur in these 3 domains in children with JIA (7,24). All item banks measure their domain-specific levels of functioning accurately across a wide range of level of functioning and in the clinically most relevant direction from the mean. The PROMIS short forms and CATs provided reliable estimations for the majority of patients. CATs outperformed short forms in terms of test length and number of reliably estimated patients.

The aim of the pediatric Dutch-Flemish PROMIS group is to improve the measurements of patient-reported outcomes in The Netherlands and Belgium by providing researchers and health care professionals access to the generic pediatric PROMIS item

banks, short forms, and CATs. The current study supports the application of CATs in clinical samples. The PROMIS item banks outperformed legacy instruments (the PedsQL) by providing more reliable measurements across a broader range of functioning.

A limitation of this study is that our sample was small and contained a large proportion of patients with inactive disease. Due to a combination of relatively good health and a small sample size, the physical function item banks did not have enough variation in responses to provide reliable parameter estimates; particularly, 2 items from the upper extremity item banks had outlying discrimination parameters due to a lack of variety of item responses. Due to the skewed data, a moderate ceiling effect was present for the mobility and upper extremity item banks. This might indicate that there are not enough items with a high difficulty present in these item banks to discriminate between patients with healthier levels of functioning. However, having fewer precise measurements at a healthy level of functioning is less important than having precise measurements in the clinical

Table 4. Reliability and test-retest reliability of measurements of the full-length (FL) item banks, short forms (SF), and computerized adaptive testing (CAT) of the pediatric PROMIS item banks in a sample of children with juvenile idiopathic arthritis (n = 155)*

PROMIS item	Mean FL SE(θ)	Mean SF SE(θ)	Mean CAT SE(θ)	Mean CAT SE(θ)	Mean CAT items administered	FL, no. of items	SF, no. of items	FL, ICC (95% CI) (n = 101)	SF, ICC (n = 101)	SDC (n = 101)
Anger scale	0.37 (57.4)	0.51 (57.4)	0.40 (57.4)	0.40 (57.4)	3.6	5	5	0.70 (0.59-0.70)	0.70	15.3
Anxiety	0.36 (55.5)	0.52 (47.1)	0.41 (54.2)	0.41 (54.2)	5.6	13	8	0.77 (0.68-0.84)	0.76	13.5
Depressive symptoms	0.34 (57.4)	0.75 (50.3)	0.40 (56.8)	0.40 (56.8)	5.2	13	8	0.79 (0.70-0.85)	0.77	14.2
Fatigue	0.20 (79.4)	0.40 (69.7)	0.31 (73.5)	0.31 (73.5)	4.7	23	10	0.87 (0.82-0.91)	0.85	17.2
Mobility	0.30 (67.5)	0.54 (50.3)	0.37 (63.1)	0.37 (63.1)	5.3	23	8	0.84 (0.76-0.89)	0.81	13.3
Pain interference	0.27 (69.7)	0.40 (66.5)	0.36 (68.4)	0.36 (68.4)	4.5	13	8	0.83 (0.77-0.89)	0.82	13.6
Peer relationships	0.29 (72.3)	0.41 (52.9)	0.36 (62.6)	0.36 (62.6)	5.8	15	8	0.69 (0.58-0.78)	0.72	18.7
Upper extremity	0.38 (48.7)	0.84 (41.7)	0.45 (44.9)	0.45 (44.9)	6.0	29	8	0.86 (0.80-0.90)	0.84	12.1

* PROMIS = Patient-Reported Outcomes Measurement Information System; SE(θ) = SE of theta; ICC = intraclass correlation coefficient; 95% CI = 95% confidence interval; SDC = smallest detectable change.

† Number of patients with an SE(θ) <0.32. An SE(θ) of 0.32 equals a reliability of 0.90.

range. The skewness of the data also has an effect on the informative value of items, and consequently, on the SE(θ). The item banks peer relationships, mobility, and upper extremity displayed lower item thresholds and some local dependent item pairs, also due to skewness. Similar skewed data were found in a US sample of patients with JIA (8). Despite the small sample size, this study shows strong psychometric properties in this population.

The psychometric properties of the PROMIS item banks in this study were similar to the properties reported in the developmental phase of the instruments (17–22) in terms of IRT model and item fit. For the study of US children with JIA, fit indices were not available. Brandon et al (8) investigated the discriminative validity across different levels of disease activity in children with JIA. Their study found discriminative abilities for the fatigue, mobility, pain interference, and upper extremity item banks. Our findings support these results, except for the fatigue item bank. This is possibly due to different methods of determining disease activity. In this study, a comparison was only made between presence and total absence of disease activity, as there were only limited retrospective data available to assess disease activity, and few patients with disease activity to facilitate group comparisons. The reliability of the measurements of the Dutch JIA sample were generally higher than those found in the US sample (8). This is possibly due to differences in model calibration and parameterization. To our knowledge, no studies have been published that assess the test–retest reliability of the full pediatric item banks. In the current study, test–retest reliability was sufficient for all item banks, except the peer relationships item bank (ICC 0.69). Varni et al (43) assessed the test–retest reliability of the pediatric short forms and found similar results. Additionally, the current study displayed similar test–retest reliability for short forms and full-length item banks.

To enable international comparisons of PROMIS T scores, differential item functioning (DIF) needs to be assessed between The Netherlands and the US. As the US data on JIA children were unobtainable, assessing DIF was not possible in this study. A next step is to calibrate the pediatric item banks in a normative Dutch sample and perform DIF analyses with the US normative sample. In conclusion, the current study demonstrates sufficient psychometric properties for the pediatric PROMIS item banks in children with JIA and provides evidence for the advantages of using the PROMIS CATs in Dutch clinical populations.

AUTHOR CONTRIBUTIONS

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

Study conception and design. Terwee, van Oers, Joosten, Grootenhuus, van Rossum, Haverman.

Acquisition of data. Van Oers, Joosten, van den Berg, Schonenberg-Meinema, Dolman, ten Cate, Roorda, van Rossum, Haverman.

Analysis and interpretation of data. Luijten, Terwee, Haverman.

ROLE OF THE STUDY SPONSOR

Pfizer Pharmaceuticals had no role in the study design or in the collection, analysis, or interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication. Publication of this article was not contingent upon approval by Pfizer Pharmaceuticals.

REFERENCES

- Schepers SA, Sint Nicolaas SM, Haverman L, Wensing M, Schouten van Meeteren AY, Veening MA, et al. Real-world implementation of electronic patient-reported outcomes in outpatient pediatric cancer care. *Psychooncology* 2017;26:951–9.
- Haverman L, van Oers HA, Limperg PF, Hijmans CT, Schepers SA, Sint Nicolaas SM, et al. Implementation of electronic patient reported outcomes in pediatric daily clinical practice: the KLIK experience. *Clin Pract Pediatr Psychol* 2014;2:50–67.
- Ayers DC. Implementation of patient-reported outcome measures in total knee arthroplasty. *J Am Acad Orthop Surg* 2017;25 Suppl 1:S48–50.
- Haskell A, Kim T. Implementation of patient-reported outcomes measurement information system data collection in a private orthopedic surgery practice. *Foot Ankle Int* 2018;39:517–21.
- Stehlik J, Rodriguez-Correa C, Spertus JA, Biber J, Nativi-Nicolau J, Zickmund S, et al. Implementation of real-time assessment of patient-reported outcomes in a heart failure clinic: a feasibility study. *J Card Fail* 2017;23:813–6.
- Wallwiener M, Heindl F, Brucker SY, Taran FA, Hartkopf A, Overkamp F, et al. Implementation and feasibility of electronic patient-reported outcome (ePRO) data entry in the PRAEGNANT real-time advanced and metastatic breast cancer registry. *Geburtshilfe Frauenheilkd* 2017;77:870–8.
- Haverman L, Grootenhuus MA, van den Berg JM, van Veenendaal M, Dolman KM, Swart JF, et al. Predictors of health-related quality of life in children and adolescents with juvenile idiopathic arthritis: results from a web-based survey. *Arthritis Care Res (Hoboken)* 2012;64:694–703.
- Brandon TG, Becker BD, Bevans KB, Weiss PF. Patient-Reported Outcomes Measurement Information System tools for collecting patient-reported outcomes in children with juvenile arthritis. *Arthritis Care Res (Hoboken)* 2017;69:393–402.
- Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, 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 Suppl 1:S3–11.
- Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol* 2010;63:1179–94.
- 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 Suppl 1:133–41.
- Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Qual Life Res* 2010;19:125–36.
- Haverman L, Grootenhuus MA, Raat H, van Rossum MA, van Dulmen-den Broeder E, Hoppenbrouwers K, et al. Dutch-Flemish translation of nine pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). *Qual Life Res* 2016;25:761–5.
- Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, Hays RD. The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information

- System (PROMIS). *Arthritis Care Res (Hoboken)* 2011;63 Suppl 11:S486–90.
15. Witter JP. The promise of Patient-Reported Outcomes Measurement Information System turning theory into reality: a uniform approach to patient-reported outcomes across rheumatic diseases. *Rheum Dis Clin North Am* 2016;42:377–94.
 16. Petty RE, Southwood TR, Manners P, Baum J, Glass DN, Goldenberg J, et al. International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *J Rheumatol* 2004;31:390–2.
 17. Irwin DE, Stucky BD, Langer MM, Thissen D, DeWitt EM, Lai JS, et al. PROMIS Pediatric Anger Scale: an item response theory analysis. *Qual Life Res* 2012;21:697–706.
 18. Irwin DE, Stucky BD, Langer MM, Thissen D, Dewitt EM, Lai JS, et al. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Qual Life Res* 2010;19:595–607.
 19. Lai JS, Stucky BD, Thissen D, Varni JW, DeWitt EM, Irwin DE, et al. Development and psychometric properties of the PROMIS pediatric fatigue item banks. *Qual Life Res* 2013;22:2417–27.
 20. Varni JW, Stucky BD, Thissen D, Dewitt EM, Irwin DE, Lai JS, et al. PROMIS Pediatric Pain Interference Scale: an item response theory analysis of the pediatric pain item bank. *J Pain* 2010;11:1109–19.
 21. Dewitt DA, Thissen D, Stucky BD, Langer MM, DeWitt EM, Irwin DE, et al. PROMIS Pediatric Peer Relationships Scale: development of a peer relationships item bank as part of social health measurement. *Health Psychol* 2013;32:1093–103.
 22. DeWitt EM, Stucky BD, Thissen D, Irwin DE, Langer M, Varni JW, et al. Construction of the eight-item patient-reported outcomes measurement information system pediatric physical function scales: built using item response theory. *J Clin Epidemiol* 2011;64:794–804.
 23. Varni JW, Seid M, Kurtin PS. PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations. *Med Care* 2001;39:800–12.
 24. Varni JW, Seid M, Smith Knight T, Burwinkle T, Brown J, Szer IS. The PedsQL in pediatric rheumatology: reliability, validity, and responsiveness of the Pediatric Quality of Life Inventory Generic Core Scales and Rheumatology Module. *Arthritis Rheum* 2002;46:714–25.
 25. Wulfraat N, van der Net JJ, Ruperto N, Kamphuis S, Prakken BJ, Ten Cate R, et al. The Dutch version of the Childhood Health Assessment Questionnaire (CHAQ) and the Child Health Questionnaire (CHQ). *Clin Exp Rheumatol* 2001;19 Suppl 23:S111–5.
 26. Takken T, van den Eijkhof F, Hoijtink H, Helders PJ, van der Net J. Examining the psychometric characteristics of the Dutch childhood health assessment questionnaire: room for improvement? *Rheumatol Int* 2006;26:979–83.
 27. Hersh A. Measures of health-related quality of life in pediatric systemic lupus erythematosus: Childhood Health Assessment Questionnaire (C-HAQ), Child Health Questionnaire (CHQ), Pediatric Quality of Life Inventory Generic Core Module (PedsQL-GC), Pediatric Quality of Life Inventory Rheumatology Module (PedsQL-RM), and Simple Measure of Impact of Lupus Erythematosus in Youngsters (SMILEY). *Arthritis Care Res (Hoboken)* 2011;63 Suppl 11:S446–53.
 28. IBM Corporation. SPSS Statistics for Windows, version 24.0. Armonk (NY): IBM Corporation; 2016.
 29. R Core team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.
 30. Rosseel Y. Lavaan: an R package for structural equation modeling. *J Stat Softw* 2012;48:1–36.
 31. Schermelleh-Engel K, Moosbrugger H, Müller H. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods Psychol Res* 2003;8:23–74.
 32. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 2007;45 Suppl 1:S22–31.
 33. Mokken RJ. A theory and procedure of scale analysis. The Hague: Mouton; 1971.
 34. Van der Ark LA. Mokken Scale analysis in R. *J Stat Softw* 2007;20:19.
 35. Chalmers RP. Mirt: a multidimensional item response theory package for the R Environment. *J Stat Softw* 2012;48.
 36. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the Multidimensional Graded Response Model. *Front Psychol* 2016;7:109.
 37. Forero CG, Maydeu-Olivares A. Estimation of IRT graded response models: limited versus full information methods. *Psychol Methods* 2009;14:275–99.
 38. Kang T, Chen TT. Performance of the Generalized S-X² Item Fit Index for polytomous IRT models. *J Educ Meas* 2008;45:391–406.
 39. Magis D, Raïche G. CatR: an R package for computerized adaptive testing. *Appl Psychol Meas* 2011;35:576–7.
 40. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. *Appl Psychol Meas* 2009;33:419–40.
 41. Wainer H, Dorans NJ, Flaugher R, Green R, Mislevy J. Computerized adaptive testing: a primer. Mahwah (NJ): Taylor & Francis; 2000.
 42. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge (UK): Cambridge University Press; 2011.
 43. Varni JW, Magnus B, Stucky BD, Liu Y, Quinn H, Thissen D, et al. Psychometric properties of the PROMIS pediatric scales: precision, stability, and comparison of different scoring and administration options. *Qual Life Res* 2014;23:1233–43.