Qual Life Res (2014) 23:1233–1243
DOI 10.1007/s11136-013-0544-0

Psychometric properties of the PROMIS® pediatric scales: precision, stability, and comparison of different scoring and administration options

James W. Varni · Brooke Magnus · Brian D. Stucky · Yang Liu · Hally Quinn · David Thissen · Heather E. Gross · I-Chan Huang · Darren A. DeWalt

Accepted: 23 September 2013 / Published online: 2 October 2013

Abstract

Objectives The objectives of the present study are to investigate the precision of static (fixed-length) short forms versus computerized adaptive testing (CAT) administration, response pattern scoring versus summed score conversion, and test–retest reliability (stability) of the Patient-Reported Outcomes Measurement Information System (PROMIS®) pediatric self-report scales measuring the latent constructs of depressive symptoms, anxiety, anger, pain interference, peer relationships, fatigue, mobility, upper extremity functioning, and asthma impact with polytomous items.

Methods Participants ($N = 331$) between the ages of 8 and 17 were recruited from outpatient general pediatrics and subspecialty clinics. Of the 331 participants, 137 were diagnosed with asthma. Three scores based on item response theory (IRT) were computed for each respondent: CAT response pattern expected a posteriori estimates, short-form response pattern expected a posteriori estimates, and short-form summed score expected a posteriori estimates. Scores were also compared between participants with and without asthma. To examine test–retest reliability, 54 children were selected for retesting approximately 2 weeks after the first assessment.

Results A short CAT (maximum 12 items with a standard error of 0.4) was found, on average, to be less precise than the static short forms. The CAT appears to have limited usefulness over and above what can be accomplished with the existing static short forms (8–10 items). Stability of the scale scores over a 2-week period was generally supported.

Conclusion The study provides further information on the psychometric properties of the PROMIS pediatric scales and extends the previous IRT analyses to include precision estimates of dynamic versus static administration, test–retest reliability, and validity of administration across groups. Both the positive and negative aspects of using CAT versus short forms are highlighted.

Keywords PROMIS · Pediatrics · Self-report · Patient-reported outcomes · Item response theory · Computerized adaptive testing

Abbreviations

PROMIS	Patient-Reported Outcomes Measurement Information System
HRQOL	Health-related quality of life
NIH	National Institutes of Health

J. W. Varni
Department of Pediatrics, College of Medicine, Texas A&M University, College Station, TX, USA

J. W. Varni (✉)
Department of Landscape Architecture and Urban Planning, College of Architecture, Texas A&M University, 3137 TAMU, College Station, TX 77843-3137, USA
e-mail: jvarni@arch.tamu.edu

B. Magnus · Y. Liu · H. Quinn · D. Thissen
Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

B. D. Stucky
RAND Corporation, Santa Monica, CA, USA

H. E. Gross · D. A. DeWalt
Division of General Medicine and Clinical Epidemiology, Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

I.-C. Huang
Department of Health Outcomes and Policy, Institute for Child Health Policy, University of Florida, Gainesville, FL, USA

Introduction

The Patient-Reported Outcomes Measurement Information System (PROMIS®) is a National Institutes of Health (NIH) initiative created to advance the assessment of patient-reported outcomes (PROs) in chronic diseases. Items are evaluated using item response theory (IRT) to derive item banks with scores that are theoretically reliable and valid along the full spectrum of the latent trait [1]. A primary objective is to develop item banks and computerized adaptive testing (CAT) potentially applicable across a variety of chronic disorders [2]. An additional objective has been to develop multiple unidimensional static (fixed-length) short forms in addition to dynamic CAT administration of the item banks.

During the past 10 years, the PROMIS Pediatric Cooperative Group has developed pediatric self-report item banks with polytomous items for ages 8–17 years across five generic health domains (physical functioning, pain, fatigue, emotional health, and social health) consistent with the larger PROMIS network [3]. It was anticipated that measures of these five generic health domains would be applicable across pediatric chronic health conditions, so generic or non-disease-specific scales were developed [4–10]. These five generic health domains have thus far been further delineated into the eight latent constructs of depressive symptoms, anxiety, anger, pain interference, peer relationships, fatigue, mobility, and upper extremity functioning [4–10]. An asthma-specific measure has also been created [11, 12].

The items were initially developed through an extensive review of the literature, expert review, and qualitative methods (focus groups and cognitive interviewing) [13, 14]. Subsequent quantitative methods utilized IRT procedures to develop item banks on a common metric with polytomous items, minimizing local dependence of items and differential item functioning, in addition to creating unidimensional static (fixed-length) short forms of the latent constructs [4–11]. All of the static short forms created consist of 8 items, except for fatigue (10 items) and anger (6 items). However, to date, the precision of different administration and scoring methods (static short forms versus CAT) and test–retest reliability (stability) have not been reported for these PROMIS pediatric self-report scales.

Consequently, the objectives of the present study are to investigate the precision of different scoring and administration options, including the measurement properties of static short forms versus dynamic CAT administration of the item banks and response pattern scoring versus summed score conversion. We also evaluated the test–retest reliability of these recently developed PROMIS pediatric scales measuring the nine latent constructs of depressive symptoms, anxiety, anger, pain interference, peer relationships, fatigue, mobility, upper extremity functioning, and asthma impact.

Method

Data collection

Participants ($N = 331$) were recruited between September 2009 and March 2010 from outpatient general pediatrics and subspecialty clinics of two public universities in North Carolina ($N = 267$) and Texas ($N = 64$). To be eligible for the study, participants had to be between the ages of 8 and 17; able to speak and read English; and have the ability to interact with a computer screen, keyboard, and mouse. Children were excluded from the study if the researcher determined they had a medical or psychiatric condition that precluded participation or a cognitive or other impairment that would interfere with completing the survey. Children with asthma comprised approximately 40 % of the sample ($N = 137$). Children's asthma status was self-reported by their parents or caregivers. To be considered, children had to receive a physician diagnosis of asthma and to be using asthma medication at the time of the study.

Research assistants approached parents of children between the ages of 8 and 17. In addition, informational recruitment fliers, brochures, and a study poster were placed in clinic waiting rooms. Researchers provided a brief explanation of the study, and if the participant was willing and eligible to participate, administered consent and assent forms.

All participants completed the majority of the survey on a computer. Surveys were usually completed before and/or after the child's clinic visit; however, an appointment could also be scheduled for another time if the participant preferred. Parents were asked to answer a few demographic items and for the asthma sample, questions about their children's asthma status. Children with asthma completed the Asthma Control Test [15, 16] and the Pediatric Asthma Quality of Life Questionnaire [17] on paper while their parent answered the computer-based questions. Then, the child completed the remainder of the survey (i.e., all PROMIS items) on the computer. It took participants approximately 15–30 min to complete the survey, and they received a gift card for participation.

To examine test–retest reliability (stability), a subset of the children ($N = 130$) were invited for retesting. Participation in the follow-up study was not required, and a target sample size of approximately 50 was set, after which recruitment was discontinued. Fifty-four children completed a second assessment (Time 2) approximately 2 weeks after the first assessment (Time 1). Children were ineligible to participate in the follow-up if they were experiencing an acute illness at enrollment. Research assistants contacted parents of eligible children by phone to initiate the follow-up. If children were not sick at the time of the follow-up phone call, they were asked to complete the online survey by accessing it from their home computers.

Domains and items

All children responded to items from eight of the PROMIS pediatric scales: depressive symptoms, anxiety, anger, pain interference, peer relationships, fatigue, mobility, and upper extremity. In addition, the asthma impact scale was administered to children with asthma. All items used a 7-day recall period and one of the two sets of standardized 5-point response options: *never*, *almost never*, *sometimes*, *often*, and *almost always*, for all scales except the physical functioning scales (mobility and upper extremity); *with no trouble*, *with a little trouble*, *with some trouble*, *with a lot of trouble*, *not able to do* for the latter scales. The scoring directions of all scales are suggested by their names; higher scores on peer relations, mobility, and upper extremity indicate better functioning, whereas higher scores on depressive symptoms, anxiety, anger, pain interference, fatigue, and asthma indicate poorer functioning.

Administration and scoring

Because one goal of this study was to compare the results obtained with fixed or static short forms with results from CAT administration, item administration was arranged so that the participants received both short form and CAT scores. This was accomplished by administering the CAT first; then any items on the recommended short forms that had not been included in the child's adaptive test were administered as well. Thus, for each measure, the participants first answered the items that were administered using CAT, and then they were also administered all of the items on the previously published static short forms. If they were administered one of the short-form items as part of the CAT, they were not administered this item again, but their response to this item was used to create both their CAT and short-form scores. All short forms contained 8 items, except fatigue (10) and anger (6), in the published static short forms [4–11].

The CAT was administered using the PROMIS Assessment Center software [18], with maximum posterior weighted information item selection, and stopped when the posterior standard deviation dropped below 0.4 standard units; the CAT administered a minimum of 5 and a maximum of 12 items. Short forms were those recommended in the original development of the scales [4–11]. There was no CAT for the anger scale, because that bank only has six items, all of which were administered as its short form [8]. The second test administration for the retest sample had the same structure as the first; the CAT was completed first, followed by items to complete the short form for each domain.

Three scores based on IRT were computed for each respondent: CAT response pattern expected a posteriori (CAT**u**EAP) estimates, short-form response pattern expected a posteriori (SF**u**EAP) estimates, and short-form summed score expected a posteriori (SF**x**EAP) estimates; all are estimates of the latent domain score (see [19] for the IRT scoring algorithms).

Statistical and psychometric analysis plan

The precision of the three types of scores was examined using the posterior standard deviations that are reported as the standard errors of measurement of the scores and root mean square errors (RMSE), the square root of the average error variance across all participants. These values were examined graphically to determine how the precision varied at different levels of the latent construct.

To check the extent to which the CAT system administered fewer or more items than are on the short forms, the distribution of CAT test lengths was summarized with descriptive statistics. The overlap of item administration between the CAT system and the short forms was investigated by computing the proportions of CAT items that were also short-form items, and the proportion of short-form items administered by the CAT system.

Test–retest reliability was assessed by computing the correlations between scores from Time 1 and Time 2 for all three scoring methods for the nine PROMIS pediatric domains. For comparison, simulation was used to compute an IRT analog to classical internal consistency reliability. There are several approximate methods to estimate reliability for scales built with IRT [20]; however, a method more straightforward than any formula is to simulate item response data and compute the squared correlation of the IRT score estimates with the generating values of the underlying score. That was done for these scales, using simulation sample sizes of 30,000.

Correlations between CAT and short-form scores were computed for the Time 1 forms to investigate their comparability. Relationships among the nine scales were examined by computing correlation coefficients among the scores. Group differences analyses using *t* statistics for the comparison of children diagnosed with asthma versus the other children were computed as standardized effect size estimates, for the three scoring methods, to investigate

Table 1 Participant demographics

Child demographics	Full sample	Children with asthma	Children without asthma	Comparison of groups with and without asthma	
	$N = 331$ n (%)	$N = 137$ n (%)	$N = 194$ n (%)	t statistic	Sig[c]
Child's gender					
Female	170 (51.4)	65 (47.4)	105 (54.1)	1.20	NS
Child's age (years)					
8–12	184 (55.6)	81 (59.1)	103 (53.1)		
13–17	147 (44.4)	56 (40.9)	91 (46.9)		
Age (*M*, SD)	12.13 (2.62)	12.04 (2.62)	12.19 (2.62)	0.50	NS
Child's race					
White or Caucasian	162 (48.9)	58 (42.3)	104 (53.6)	2.03	0.04
Black or African-American	130 (39.3)	63 (46.0)	67 (34.5)	2.09	0.04
Asian	8 (2.4)	2 (1.5)	6 (3.1)	0.95	NS
Multiple races	5 (1.5)	2 (1.5)	3 (1.5)	0.06	NS
Other	22 (6.6)	9 (6.6)	13 (6.7)	0.05	NS
Not provided	4 (1.2)	3 (2.2)	1 (.5)		
Child's ethnicity					
Hispanic	35 (10.6)	21 (15.3)	14 (7.2)	2.25	0.03
Not provided	22 (6.6)	12 (8.8)	10 (5.2)		
Child's history of other health problems					
None	146 (44.1)	55 (40.1)	91 (46.9)		
1 Health problem	119 (36.0)	47 (34.3)	72 (37.1)		
≥2 Health problem	64 (19.3)	33 (24.1)	31 (16.0)		
Missing	2 (0.6)	2 (1.5)	0		
Number of other health problems (*M*, SD)	0.85 (0.99)	0.99 (1.14)	0.75 (.87)	2.21	0.03
Most common other health problems[a]					
ADHD	73 (22.1)	34 (24.8)	39 (20.1)	0.94	NS
Epilepsy or other seizure disorder	14 (4.2)	5 (3.6)	9 (4.6)	0.43	NS
Intestinal disorder	15 (4.5)	4 (2.9)	11 (5.7)	1.26	NS
Overweight	46 (13.9)	24 (17.5)	22 (11.3)	1.50	NS
Premature birth	34 (10.3)	17 (12.4)	17 (8.8)	1.19	NS
Mental health disorders	33 (10.0)	16 (11.7)	17 (8.8)	0.82	NS
Rheumatic disease	11 (3.3)	5 (3.6)	6 (3.1)	0.25	NS
Asthma sample only[b]					
Good asthma control	78 (56.9)				
Poor asthma control	59 (43.1)				
Guardian's relationship to the child					
Mother or stepmother	275 (83.1)	109 (79.6)	166 (85.6)	1.27	NS
Father or stepfather	36 (10.9)	14 (10.2)	22 (11.3)	0.30	NS
Grandparent	13 (3.9)	8 (5.8)	5 (2.6)	1.42	NS
Guardian or other	6 (1.9)	5 (3.6)	1 (0.5)	1.86	NS
Missing	1 (0.3)	1 (0.7)	0		
Guardian education level					
≤8th grade or some high school	31 (9.4)	13 (9.5)	18 (9.3)	0.07	NS
High school degree/GED	68 (20.5)	29 (21.2)	39 (20.1)	0.25	NS
Some college/technical degree	122 (36.9)	50 (36.5)	72 (37.1)	0.10	NS
College or advanced degree	108 (32.6)	44 (32.1)	64 (33.0)	0.15	NS
Missing	2 (0.6)	1 (0.7)	1 (0.5)		

NS not significant (p value was greater than 0.10)

[a] Parents reported more than 1 condition for some children; there were many other conditions reported in lower frequency (<3 %) than the conditions listed

[b] Asthma control measured by the asthma control test

[c] Independent samples t tests

whether one scoring method might be more responsive to group differences than others.

Results

Participants

A total of 331 children were recruited and administered the questionnaires at Time 1. Of these 331 children, 137 were diagnosed with asthma. Table 1 describes the background characteristics of the study participants.

Of the final study sample, a slight majority were female ($n = 170$, 51.4 %) and between the ages of 8–12 years ($n = 184$, 55.6 %), with an average age of 12.1. The largest racial group was White ($n = 162$, 48.9 %), with representative numbers from Black ($n = 130$, 39.3 %) and Hispanic ($n = 35$, 10.6 %) participants. The asthma and non-asthma groups did not differ in gender or age. However, the proportions of African-American and Hispanic participants were higher in the asthma group ($p = 0.04$ and 0.03, respectively).

The majority of guardians who completed the demographic items were the children's parents ($n = 311$, 94.0 %), with most of these being the child's mother ($n = 275$, 83.1 %). Most of the parents had either some college education ($n = 122$, 36.9 %) or a college or advanced degree ($n = 108$, 32.6 %). Majority of children ($n = 185$, 55.9 %) were described by their parents as having health problems other than or in addition to asthma. The most commonly reported other health problems were attention deficit disorder ($n = 73$, 22.1 %), being overweight ($n = 46$, 13.9 %), being born prematurely ($n = 34$, 10.3 %), and mental health disorders ($n = 33$, 10.0 %). On average, children with asthma experienced more additional health problems (0.99 compared to 0.75 for children without asthma, $p = 0.03$). However, there were no significant differences in the numbers experiencing the most common other health problems.

Score distributions

Figure 1 displays histograms of the CAT and short-form scores across all nine domains. Several results are clear from Fig. 1. First, all scoring distributions are roughly the same for the three types of scores within each domain. The two types of scores based on the short forms (response pattern EAPs and summed score EAPs) are most similar; this is to be expected, because these scores are computed from the same responses. Second, all domains show either a floor or ceiling effect. Domains in which higher scores indicate higher functioning (peer relations, mobility, and upper extremity) display a ceiling, whereas domains in which higher scores indicate lower functioning have a floor. Third, the adaptive nature of the CAT is evident from the histograms. The score range tends to be greater for the CAT than the short forms; this is because an adaptive test is better able to measure the functioning of people falling at the low or high ends of the scale. Using the short forms, frequencies are more likely to accumulate at the maximum (or minimum) scores, whereas using the CAT allows for more nuanced measurement at the extreme ends of the scale.

Short form versus CAT scores

Correlations between CAT and the fixed-length short-form scores were computed from the Time 1 data (see Table 2). All correlations were very high. As expected, the short-form scores were more highly correlated with each other than they were with the CAT scores, but all correlations exceed 0.93.

Precision of measurement

Figure 2 shows the IRT standard errors for each of the three types of scores plotted against the scores. RMSE (the square root of the average squared standard errors) is shown in the lower right hand corner of each plot. RMSE for summed score EAPs is only slightly greater than that for the response pattern EAPs, indicating that minimal precision is lost in using the summed score EAPs for the short form.

In the middle range of the latent constructs, the standard errors for the CAT scores are greater than those for the summed score and response pattern EAPs for the short forms. This is due to the stopping rule of the CAT, which meant that very few items were administered to people falling in this mid-range. However, at the extremes of some scales, the CAT standard errors drop below the short-form scores, indicating that the CAT has adapted to the participant's level on the latent construct and provided slightly more precise measurement for people scoring at the high or low ends.

Items administered by the CAT system

Table 3 describes the item selection of the CATs relative to the short forms in more detail. The first four columns of Table 3 summarize the numbers of items administered by the CAT system. The minimum number of items was usually 5, as prescribed (there was one participant who only completed 4 items, for pain interference), and the

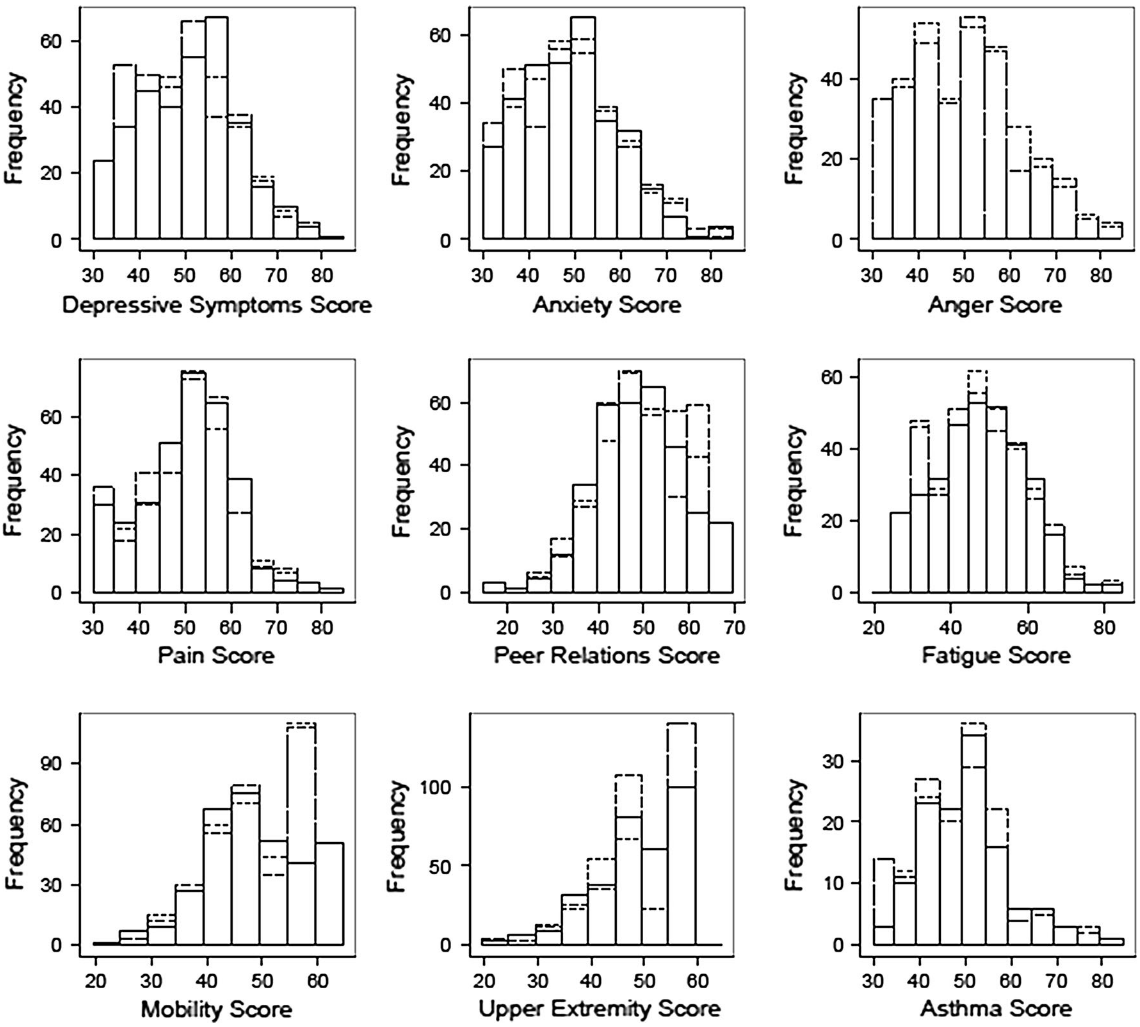

Fig. 1 Histograms from CAT response pattern EAP scores (*solid outline*), short-form response pattern EAP scores (*outline with short dashes*), and short-form summed score EAP scores (*outline with long dashes*)

Table 2 Correlations between CAT scores and (alternative) short-form scores

Scale	CAT**u**EAP—SF**u**EAP	CAT**u**EAP—SF**x**EAP	SF**u**EAP—SF**x**EAP
Depressive symptoms	0.98	0.98	1.00
Anxiety	0.98	0.98	1.00
Anger	–	–	0.99
Pain interference	0.98	0.97	1.00
Peer relationships	0.95	0.95	0.99
Fatigue	0.98	0.97	1.00
Upper extremity	0.95	0.93	0.99
Mobility	0.95	0.94	0.99
Asthma impact	0.98	0.96	0.99

maximum was always the upper limit, 12 items. The range of average CAT length was 6.2–10.7 items, so on average the CAT administered about the same number of items as are on the short forms (8, except for the 10-item fatigue short form).

However, for most scales, the distribution of CAT lengths was distinctly bimodal, as shown in columns five and six of Table 3 that list the proportions of CATs that were 5 or 12 items long; for all scales except fatigue, CATs of minimum and maximum length account for the majority of the administrations. Either the CAT system achieved the precision of a standard error lower than 4 points on the *T* score scale in 5 items and stopped, or it administered the maximum number of items (12) without obtaining a score of that level of precision.

The next eight columns of Table 3 show the minimum, maximum, mean, and standard deviation of the proportion of CAT-administered items that were also on the short forms, and then the proportion of short-form items administered by the CAT system. The former, on average, ranges from 0.67 to 0.93, meaning that most CAT-

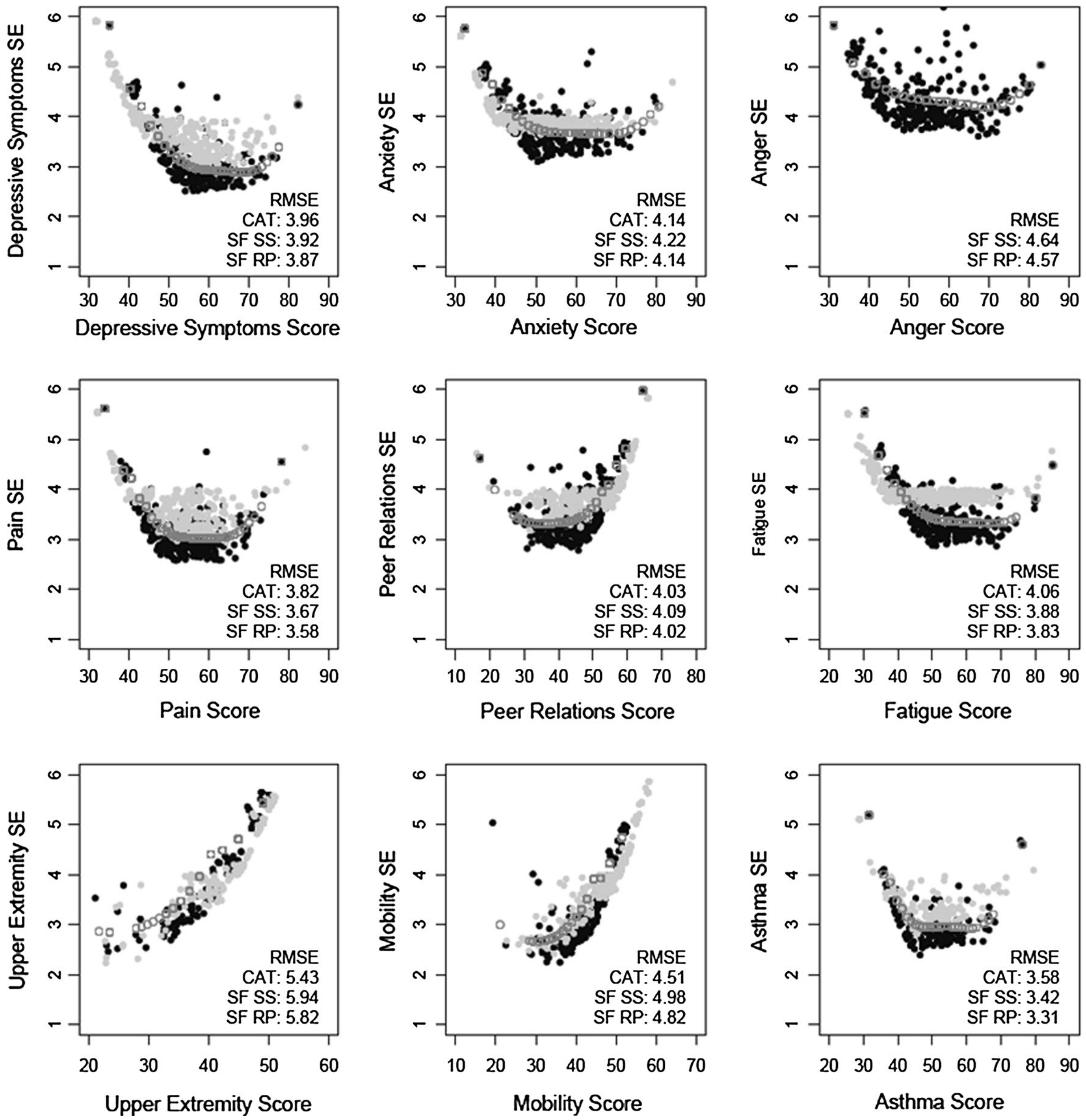

Fig. 2 IRT standard errors across scores, and RMSE, for CAT response pattern EAP scores (*light gray dots*), short-form response pattern EAP scores (*black dots*), and short-form summed score EAP scores (*gray open circles*)

administered items were short-form items; and the latter, on average, ranges from 0.66 to 0.88, meaning that the CAT system often administered a fraction of the short form.

The rightmost column of Table 3 lists the proportions of CAT administrations that were exactly subsets of the short forms. That happened often, more than half of the administrations for depressive symptoms, pain interference, fatigue, and asthma impact. When the CAT is a subset of the short form, the CAT score is necessarily less precise than the short-form score; this explains why the RMSE values for CAT scores for those scales are larger than those for the short forms.

At the other extreme, for upper extremity functioning, the CAT was rarely a subset of the short form (0.07 of the administrations), but it most often administered the maximum number of items (0.74 of the administrations). So for upper extremity, the CAT provided more precise measurement, but it did so by becoming (selectively) longer than the short form to measure in the higher score range (see Fig. 1). The mobility scale behaved similarly, but not so extremely. That is, for mobility, the CAT was slightly longer than the short form and administered the maximum number of items to a fairly high proportion of participants (0.42) compared to other domains, but not as high as the upper extremity scale.

Table 3 Numbers of items administered by the CAT system and overlap with the short forms (SF)

Scale	CAT length				Prop. CATS of length		Proportion of CAT items that were also on the SF				Proportion of SF items administered by the CAT				Subset
	Min.	Max.	*M*	SD	5	12	Min.	Max.	*M*	SD	Min.	Max.	*M*	SD	
Depressive symptoms	5	12	7.1	2.9	0.54	0.23	0.58	1.0	0.89	0.17	0.63	1.0	0.74	0.13	0.69
Anxiety	5	12	8.1	2.7	0.24	0.26	0.56	1.0	0.84	0.14	0.50	1.0	0.82	0.18	0.38
Pain interference	4	12	6.4	2.6	0.71	0.16	0.50	1.0	0.93	0.13	0.50	1.0	0.71	0.15	0.78
Peer relationships	5	12	7.6	2.9	0.43	0.23	0.50	1.0	0.82	0.13	0.38	1.0	0.75	0.21	0.26
Fatigue	5	12	8.1	2.4	0.13	0.18	0.55	1.0	0.88	0.16	0.50	1.0	0.68	0.12	0.59
Upper extremity	5	12	10.7	2.5	0.09	0.74	0.20	1.0	0.67	0.13	0.13	1.0	0.88	0.18	0.07
Mobility	5	12	8.3	3.3	0.46	0.42	0.45	1.0	0.75	0.23	0.50	1.0	0.68	0.08	0.40
Asthma impact	5	12	6.2	2.5	0.76	0.14	0.55	1.0	0.92	0.15	0.50	1.0	0.66	0.10	0.76

The anxiety and peer relationships scales illustrate another pattern: The CAT system adapted to measure the lower/upper ranges of the score range more precisely, but that was more than counterbalanced by less precision for the shorter CATs in the middle of the range, so in aggregate the short form outperformed the CAT in precision.

Test–retest reliability

Fifty-four participants completed the 2-week follow-up survey. Twenty-three of the 54 retested children were from the asthma group. The time interval between completion of the first assessment and 2 week follow-up survey ranged from 11 to 17 days, with a mean of 14.9 days (SD = 2.06).

The group of children who completed the follow-up survey was compared to the children who were initially invited but did not complete the follow-up. No significant differences were found between the two groups in child gender, χ^2 (1, $N = 130$) = 2.47, $p = 0.12$; race, χ^2 (2, $N = 127$) = 2.84, $p = 0.24$; ethnicity, χ^2 (1, $N = 120$) = 1.17, $p = 0.28$; or age, $t = 1.50$, $p = 0.14$.

As a measure of test–retest reliability, correlations were computed for the short-form scores and CAT scores between Time 1 and Time 2 (see Table 4). Test–retest correlations were generally high, between 0.7 and 0.8 for the depressive symptoms, anxiety, peer relationships, fatigue, and asthma impact scales. Stability was slightly lower for the physical functioning scales (upper extremity and mobility), with correlations around 0.7, and lower yet for the pain interference and anger scores, with correlations around 0.6.

Test–retest reliability was also compared with internal consistency reliability, calculated using IRT simulation-based reliability for the two short-form scores. While test–retest measures reliability within items across time, internal consistency measures reliability across items within time. Table 4 also shows the internal consistency reliability coefficients. Internal consistency was generally higher than test–retest reliability, with the exception of upper extremity functioning for which test–retest reliability was higher.

Domain correlations

Correlations across scales were calculated using the short-form response pattern scores (see Table 5).

Group difference

Table 6 shows the standardized effect size estimates and *t* statistics for the comparison of children diagnosed with asthma versus the non-asthma groups, for the three scoring methods. For this comparison, the only significant difference between the groups is for the mobility score; for the SF**u**EAPs, the average mobility scores were 46.9 (SD = 8.3) for the asthma group and 50.0 ($D = 8.5$) for the others. The effect sizes are generally similar for all of the between-group comparisons, again indicating that the scoring method makes little difference.

Discussion

This study demonstrates that scores on the PROMIS pediatric measures are highly correlated regardless of the scoring or administration technique (CAT versus static short forms, response pattern versus summed score conversion). It is especially notable that the correlations between response pattern scores and summed score conversions are either 0.99 or round to 1.00 to two decimal places, indicating that very little loss of information occurs when the convenient summed score calculations are used. This facilitates the ability to score these PROMIS pediatric short-form scales administered on paper or in systems other than Assessment Center since IRT software is not necessary for scoring these scales. It is important to note, however, that there was item overlap between the static and

Table 4 Internal consistency and test–retest reliability

Scale	Internal consistency		Test–retest reliability		
	SF**u**EAP	SF**x**EAP	SF**x**EAP	SF**u**EAP	CAT**u**EAP
Depressive symptoms	0.86	0.85	0.76	0.76	0.77
Anxiety	0.84	0.83	0.75	0.74	0.74
Anger	0.81	0.79	0.64	0.54	–
Pain interference	0.88	0.87	0.62	0.66	0.65
Peer relationships	0.84	0.83	0.76	0.81	0.67
Fatigue	0.87	0.87	0.76	0.76	0.80
Upper extremity	0.63	0.62	0.66	0.71	0.71
Mobility	0.74	0.73	0.77	0.73	0.72
Asthma impact	0.90	0.89	0.82	0.81	0.76

Table 5 Scale intercorrelations computed using short-form response pattern scores

Scale	Dep. symp.	Anxiety	Anger	Pain Int.	Peer Rel.	Fatigue	Upper extrem.	Mobility	Asthma impact
Depressive symptoms	1.00								
Anxiety	0.72	1.00							
Anger	0.71	0.70	1.00						
Pain interference	0.47	0.49	0.51	1.00					
Peer relationships	−0.35	−0.28	−0.24	−0.20	1.00				
Fatigue	0.57	0.57	0.57	0.59	−0.19	1.00			
Upper extremity	−0.37	−0.39	−0.27	−0.44	0.19	−0.32	1.00		
Mobility	−0.44	−0.42	−0.40	−0.58	0.23	−0.57	0.44	1.00	
Asthma impact	0.41	0.58	0.47	0.61	−0.12	0.68	−0.35	−0.60	1.00

Table 6 Standardized effect size estimates (*Eff.*) and *t* statistics for the comparison of children diagnosed with asthma versus the others, for the three scoring methods

Scale	SF**x**EAP			SF**u**EAP			CAT**u**EAP		
	Eff.	*t*	*p*	Eff	*t*	*p*	Eff	*t*	*p*
Depressive symptoms	−0.04	−0.34	0.74	−0.04	−0.33	0.74	−0.03	−0.30	0.77
Anxiety	0.01	0.07	0.94	0.02	0.16	0.87	−0.04	−0.31	0.76
Anger	−0.07	−0.58	0.56	−0.07	−0.62	0.53	–	–	–
Pain interference	−0.07	−0.61	0.54	−0.06	−0.6	0.55	−0.04	−0.42	0.67
Peer relationships	−0.08	−0.72	0.48	−0.07	−0.68	0.50	−0.02	−0.23	0.82
Fatigue	−0.10	−0.85	0.40	−0.14	−1.31	0.19	−10	−0.90	0.37
Upper extremity	−0.06	−0.50	0.62	−0.10	−0.79	0.43	−0.06	−0.52	0.60
Mobility	0.38	3.41	<0.01	0.37	3.30	<0.01	0.34	3.00	<0.01

Anger has no CAT

adaptive forms. Items were not administered more than once to a child; the child's response to an item on the CAT was used in computing the short-form scores. Therefore, high correlation between the two static forms and the CAT is to be expected.

The patterns of measurement precision of the nine scales were very much as expected from the IRT analyses used in the original construction of the scales [4–11]. All of the measures, to a greater or lesser extent, provide more precise measurement on the less-healthy side of the latent continua they measure. It should be noted that for measurement instruments intended for use with patients with chronic health conditions, less precision at the healthy or less severe end of the latent trait spectrum may not be problematic [21].

This study was conducted to determine how well measures would perform if the CAT was designed to be short. We found that if we shortened the CAT to a 12 item maximum with a standard error of 0.4, across most domains it was not as precise as the static short form. These CAT parameters had limited usefulness over and above what was accomplished with the static short forms (8–10 items in this case). For the physical functioning scales, upper extremity and mobility, the CAT system provided greater precision compared to short-form scores (as measured in aggregate by RMSE), but at the expense of longer tests (more items than the short forms). This result occurred because the precision associated with a CAT stopping criterion of a standard error of 4 is greater than could be achieved given the average levels of precision for the short forms. For the other scales, the short-form scores outperformed the CAT system on the average, although the CAT provided slightly more precise measurement at one end of the scale or the other. It should be noted that it is often the case that CAT administration outperforms static short forms in adults [21, 22]. For example, in adults, polytomous item CAT has demonstrated better precision than static short forms in measuring fatigue, but the short forms also demonstrated good precision for most participants [23] and have been found in a measure of depressive symptoms to perform only marginally worse than CAT [22]. Whether CAT would have performed better than the static short forms if a greater number of items had been dynamically administered was not the objective of the present study since we were primarily concerned with reducing participant burden by testing a short CAT versus static short forms. It is quite possible that CAT would have performed equally to or better than the static short forms if the CAT parameters were adjusted to administer more items (or a stopping rule at a more precise level). In this study, we were unable to test this hypothesis. Future research should investigate how the CAT algorithm might be improved. For example, the SE cut point could be lowered, which most likely would result in the administration of more items in the middle range of the latent trait and increase precision. Additionally, the total number of items administered at the extremes of the measurement continuum could be lowered in order to reduce respondent burden for some populations.

Although CAT administration has several notable hypothesized advantages over static measurement instruments, such as the potential for brief scales which may reduce respondent burden and the possibility for greater precision over the full range of the latent trait continuum, unidimensional short forms also have several practical advantages. For example, with paper and pencil administration of short forms, the need for computer access for participants is eliminated, and thus, the available administration options may be increased for some pediatric populations. On the other hand, electronic administration is often preferred since it potentially reduces data entry errors and youth are quite comfortable with computers and other mobile devices [24]. In this case, both electronic administration of static short forms and CAT administration have practical advantages over paper and pencil administration.

The stability of the scale scores over a 2-week period was very much in line with expectation, that is, there was some variability in stability; with stability coefficients lower than internal consistency reliability for all scales, except for the upper extremity scale. The scale scores were moderately correlated for the most part and show an expected pattern of relationships with patients with asthma.

There are several limitations of this study, which suggest future directions for the next generation of PROMIS pediatric scale development. Specifically in this study, the minimum number of items selected for the polytomous CAT was 5 with a maximum of 12 items. This was a decision made at the initiation of the present study, and these analyses are thus limited by this earlier decision. On average, the short forms outperformed the CAT because of how the standard error stopping rule was set (at 0.4). While this stopping rule increased the precision of the CAT, it also effectively increased the minimum number of items administered. Because the short forms have standard errors lower than 0.4 for much of the measured continua, the CAT was stopped before getting to a precision as good as the short forms. Future studies should explore using different cut points for the standard error stopping rule and changing the minimum and maximum number of items CAT administered when comparing static short forms to CAT administration, as well as methods for the administration of polytomous items such as a two-stage semi-adaptive testing strategy [22]. A further potential limitation was that not all participants selected for the test–retest reliability phase completed the second administration, which may limit the generalizability of the stability findings.

In conclusion, this study provides further support for the psychometric properties of the PROMIS pediatric scales and extends the previous IRT analyses [4–11] to include the additional measurement properties of precision estimates of dynamic versus static administration, test–retest reliability, and validity of administration across groups.

Acknowledgments This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U01AR052181. Information on the Patient-Reported Outcomes Measurement Information System (PROMIS®) can be found at http://nihroadmap.nih.gov/ and http://www.nihPROMIS.org.

References

1. Ader, D. N. (2007). Developing the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(Suppl 1), S1–S2.
2. 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-Report Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(Suppl 1), S22–S31.

3. 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 2 years. *Medical Care, 45*(Suppl 1), S3–S11.
4. Irwin, D. E., Stucky, B. D., Thissen, D., DeWitt, E. M., Lai, J. S., Yeatts, K., et al. (2010). Sampling plan and patient characteristics of the PROMIS pediatrics large-scale survey. *Quality of Life Research, 19*, 585–594.
5. Irwin, D. E., Stucky, B. D., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., et al. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research, 19*, 595–607.
6. Varni, J. W., Stucky, B. D., Thissen, D., DeWitt, E. M., Irwin, D. E., Lai, J. S., et al. (2010). PROMIS Pediatric Pain Interference Scale: An item response theory analysis of the pediatric pain item bank. *Journal of Pain, 11*, 1109–1119.
7. DeWitt, E. M., Stucky, B. D., Thissen, D., Irwin, D. E., Langer, M., Varni, J. W., et al. (2011). Construction of the eight-item patient-reported outcomes measurement information system pediatric physical function scales: Built using item response theory. *Journal of Clinical Epidemiology, 64*, 794–804.
8. Irwin, D. E., Stucky, B. D., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., et al. (2012). PROMIS Pediatric Anger Scale: An item response theory analysis. *Quality of Life Research, 21*, 697–706.
9. DeWalt, D. A., Thissen, D., Stucky, B. D., Langer, M. M., DeWitt, E. M., Irwin, D. E., Lai, J. S., Yeatts, K. B., Gross, H. E., Taylor, O., & Varni, J. W. PROMIS pediatric peer relationships scale: Development of a peer relationships item bank as part of social health measurement. *Health Psychology* (in press).
10. Lai, J.-S., Stucky, B. D., Thissen, D., Varni, J. W., DeWitt, E. M., Irwin, D. E., Yeatts, K. B., & Dewalt, D. A. Development and psychometric properties of the PROMIS® pediatric fatigue item banks. *Quality of Life Research*. doi:10.1007/s11136-013-0357-1.
11. Yeatts, K., Stucky, B. D., Thissen, D., Irwin, D. E., Varni, J. W., DeWitt, E. M., et al. (2010). Construction of the Pediatric Asthma Impact Scale (PAIS) for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Asthma, 47*, 295–302.
12. Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL^{TM} 3.0 Asthma Module to obtain scores comparable with those of the PROMIS Pediatric Asthma Impact Scale (PAIS). *Quality of Life Research, 20*, 1497–1505.
13. Walsh, T. R., Irwin, D. E., Meier, A., Varni, J. W., & DeWalt, D. A. (2008). The use of focus groups in the development of the PROMIS pediatrics item bank. *Quality of Life Research, 17*, 725–735.
14. Irwin, D. E., Varni, J. W., Yeatts, K., & DeWalt, D. A. (2009). Cognitive interviewing methodology in the development of a pediatric item bank: a patient reported outcomes measurement information system (PROMIS) study. *Health and Quality of Life Outcomes, 7*(3), 1–10.
15. Nathan, R. A., Sorkness, C. A., Kosinski, M., Schatz, M., Li, J. T., Marcus, P., et al. (2004). Development of the asthma control test: A survey for assessing asthma control. *Journal of Allergy and Clinical Immunology, 113*, 59–65.
16. Liu, A. H., Zeiger, R., Sorkness, C., Mahr, T., Ostrom, N., Burgess, S., et al. (2007). Development and cross-sectional validation of the childhood asthma control test. *Journal of Allergy and Clinical Immunology, 119*, 817–825.
17. Juniper, E. F., Guyatt, G. H., Feeny, D. H., Ferrie, P. J., Griffith, L. E., & Townsend, M. (1996). Measuring quality of life in children with asthma. *Quality of Life Research, 5*, 35–46.
18. Cella, D., Gershon, R., Bass, M., & Rothrock, N. (2012). Assessment center user manual, version 8.7. Chicago, IL: Northwestern University, Department of Medical Social Sciences.
19. Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141–186). Mahwah, NJ: Lawrence Erlbaum Associates.
20. Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika, 77*, 153–162.
21. Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*, 419–440.
22. Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*, 125–136.
23. Lai, J. S., Cella, D., Choi, S. W., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation, 92*(1 Suppl), S20–S27.
24. Varni, J. W., Limbers, C. A., Burwinkle, T. M., Bryant, W. P., & Wilson, D. P. (2008). The ePedsQL^{TM} in Type 1 and Type 2 diabetes: Feasibility, reliability and validity of the Pediatric Quality of Life Inventory^{TM} internet administration. *Diabetes Care, 31*, 672–677.