Quality of Life Research (2021) 30:2683–2695
https://doi.org/10.1007/s11136-021-02852-y

Psychometric properties of the Dutch-Flemish PROMIS® pediatric item banks Anxiety and Depressive Symptoms in a general population

L. H. Klaufus[1,2] · M. A. J. Luijten[3,4] · E. Verlinden[1] · M. F. van der Wal[1] · L. Haverman[3] · P. Cuijpers[5] · M. J. M. Chinapaw[2,4] · C. B. Terwee[2,4]

Accepted: 17 April 2021 / Published online: 13 May 2021

Abstract

Purpose This study aims to validate the Dutch-Flemish PROMIS pediatric item banks v2.0 Anxiety and Depressive Symptoms, the short forms 8a, and computerized adaptive tests (CATs) in a general Dutch population and to provide reference data.

Methods Participants ($N = 2{,}893$, aged 8–18), recruited by two internet survey providers, completed both item banks. These item banks were assessed on unidimensionality, local independence, monotonicity, Graded Response Model (GRM) item fit, and differential item functioning (DIF) for gender, age group, region, ethnicity, and language. The short forms and CATs were assessed on reliability and construct validity compared to the Revised Child Anxiety and Depression Scale short version (RCADS-22) subscales. Reference scores were calculated.

Results Both item banks showed sufficient unidimensionality, local independence, monotonicity, and GRM item fit, except for three Depressive Symptoms items that showed insufficient GRM item fit. No DIF was found when using ordinal regression analyses, except for two Depressive Symptoms items that showed DIF for language; all items showed DIF for language when using IRT PRO, except for one Anxiety item. Both short forms and CATs revealed sufficient reliability for moderate and severe levels of anxiety and depression, as well as high positive correlations with corresponding RCADS-22 subscales and slightly lower correlations with non-corresponding RCADS-22 subscales.

Conclusion The Dutch-Flemish PROMIS pediatric item banks v2.0 Anxiety and Depressive Symptoms, the short forms 8a and CATs are useful to assess and monitor anxiety and depression in a general population. Reference data are presented.

Keywords Anxiety · Depression · Pediatric · PROMIS · IRT · Validation

Introduction

Anxiety and depression are highly prevalent in children and adolescents[1] and among the leading causes of youth disability worldwide [1, 2]. Prevalence rates of child anxiety

M. J. M. Chinapaw and C. B. Terwee have jointly supervised this work.

✉ L. H. Klaufus

1 Department of Epidemiology, Health Promotion, and Health Care Innovation, Public Health Service Amsterdam, Nieuwe Achtergracht 100, Amsterdam, The Netherlands

2 Department of Public and Occupational Health, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, De Boelelaan 1117, Amsterdam, The Netherlands

3 Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Child and Adolescent Psychiatry & Psychosocial Care, Amsterdam Reproduction and Development, Amsterdam Public Health Research Institute, Meibergdreef 9, Amsterdam, The Netherlands

4 Department of Epidemiology and Data Science, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, De Boelelaan 1117, Amsterdam, The Netherlands

5 Department of Clinical, Neuro and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, Van der Boechorststraat 7, Amsterdam, The Netherlands

[1] For brevity, children and adolescents are referred to as "children" in this paper.

and depression are around 6.5 and 1.3%, respectively [1]. Long-lasting episodes of child anxiety and depression predict recurrence of these disorders [3] and the development of other psychosocial problems later in life, like substance abuse or dependence, suicidal behavior, and failure to complete secondary school [4–6]. To prevent deterioration, it is critical to assess, treat, and monitor anxiety and depression in children [7, 8].

To assess and monitor anxiety and depression in children, self-report questionnaires, based on classical test theory (CTT), are often used [7, 9, 10] (e.g., the Revised Child Anxiety and Depression Scale [RCADS] [11, 12]). Although CTT questionnaires are valuable in showing the number and severity of symptoms, it assumes that all symptoms contribute equally to severity ratings of a construct, while research has shown this is not the case [13, 14]. In addition, many CTT questionnaires are relatively long, which makes them time consuming [9]. Furthermore, the qualitative meaning of scores is not always clear [15].

To advance the measurement of self-reported health, the Patient-Reported Outcomes Measurement Information System (PROMIS®) initiative in the United States of America (U.S.) developed multiple adult and pediatric item banks, which are sets of questions measuring a same construct (e.g., anxiety or depression) [16, 17]. The use of PROMIS item banks has several advantages over the use of CTT questionnaires. PROMIS item banks have the potential to measure with a higher validity and reliability, due to a careful item selection and adaptation, and the application of Item Response Theory (IRT) [16, 18–20]. IRT is a psychometric method by which items and persons are ordered on the same scale in terms of severity of the construct. Due to this ordering, item banks can be administered through computerized adaptive testing (CAT). In CAT, items are automatically selected from an item bank, based on an individual's response to a previously completed question. With CAT, fewer items are needed to obtain a reliable result than with CTT questionnaires, which need to be administered entirely [21]. When computers are unavailable, fixed-length short forms can be used consisting of e.g. four to eight items. Furthermore, PROMIS item banks are generic in nature, which makes them universally applicable in clinical and general populations. Finally, PROMIS item banks are standardized on a universal *T*-score metric where a score of 50 represents the average of the U.S. reference population with a standard deviation (*SD*) of 10, which makes it possible to interpret results of different item banks alike.

The U.S. PROMIS pediatric item banks v1.0 Anxiety and Depressive Symptoms have been validated in a diverse set of children at public schools, hospital-based outpatient general pediatrics, and subspecialty clinics [22, 23]. These item banks were translated into, among others [24], Dutch-Flemish [25] and recently updated to versions v2.0. This study aims to validate the Dutch-Flemish PROMIS pediatric item banks v2.0 Anxiety and Depressive Symptoms, the short forms 8a, and CATs in a large sample of children from the general Dutch population and to provide reference data; it adds to former research examinations on cross-cultural validity.

Methods

Participants

Participants were children aged 8–18, who lived in the Netherlands and could read Dutch. They were recruited via their parents by two internet survey providers–Kantar Public and Panel Inzicht–from January to July 2018. Figure 1 describes the sampling procedures.

Kantar Public drew a representative sample of children from the general Dutch population and an additional sample of children with a migration background in the three biggest cities in the Netherlands, since it expected an underrepresentation of these participants. Representativeness was determined per age group 8–12 and 13–18 years on the variables gender, age, household size, ethnicity, social class, and region (deviation from gold standard < 2.5% [26]). Kantar Public expected a total response rate of 32% based on previous experiences. It offered participants a gift voucher of €1.50.

Panel Inzicht approached all parents of children aged 8–18 in their panel. It expected a response rate of 10 to 20% and offered participants €0.95.

Procedure

Participants completed an online questionnaire consisting of general questions about demographics; the PROMIS pediatric item banks v2.0 Anxiety and Depressive Symptoms [22, 23]; and the Revised Child Anxiety and Depression Scale short version (RCADS-22) [11, 12, 27]. We added one question at the end of the RCADS-22 with an opposite wording (i.e., "I feel happy") to detect respondents who completed the questionnaire without paying attention to the formulation of the questions, and one question at the end of the total questionnaire to check whether respondents participated in both internet suveys. No questions could be skipped to avoid missing data.

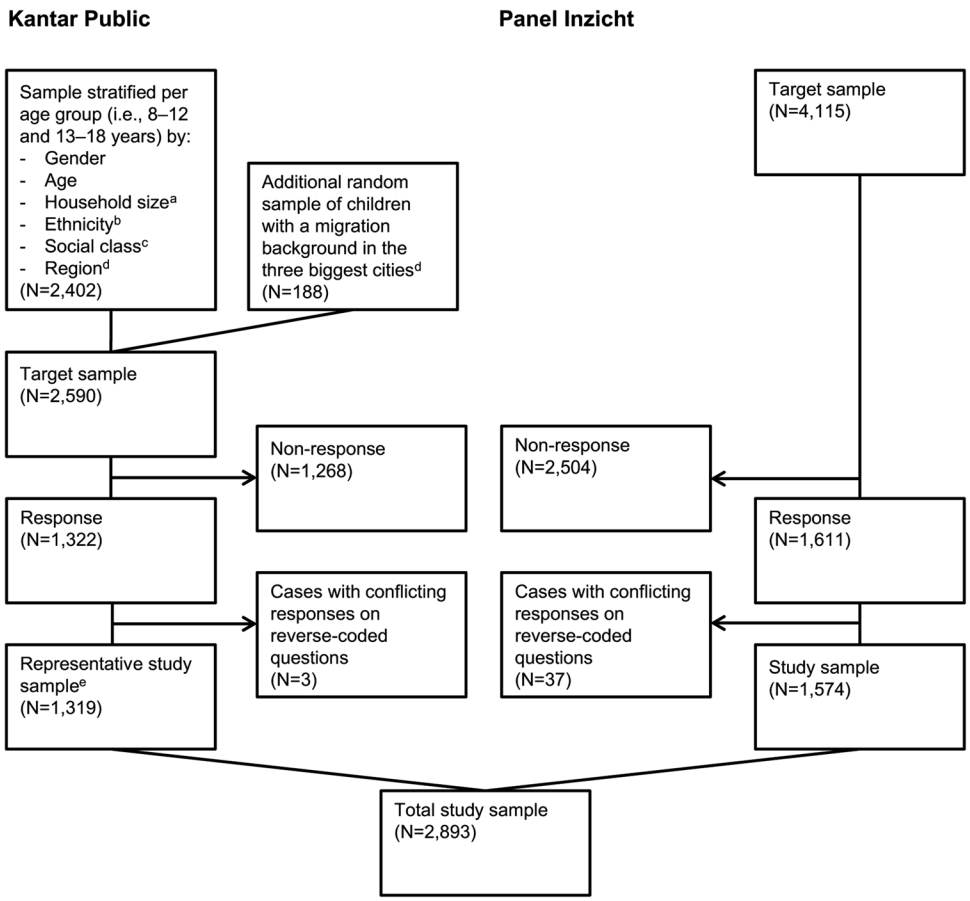

Fig. 1 Sampling procedures by two internet panel survey providers (i.e., Kantar Public and Panel Inzicht). **a** Two to six or more persons household. **b** Native, first and second generation of western immigrants (i.e., immigrants from Europe excluding Turkey, North America excluding Mexico, Oceania, Japan, and Indonesia), first and second generation of non-western immigrants (i.e., immigrants from Africa, Latin America, and Asia excluding Japan and Indonesia). **c** Five social classes. **d** Three biggest cities in the Netherlands (i.e., Amsterdam, Rotterdam, The Hague), their outskirts, region west without the three biggest cities and their outskirts, region north, region east, and region south. **e** Representative per age group 8–12 and 13–18 years old on the variables gender, age, household size, ethnicity (with the exception of native children aged 8–12), social class, and region compared to the general population in 2017

Measures

PROMIS pediatric item banks v2.0 and short forms 8a Anxiety and Depressive Symptoms [22, 23]

The PROMIS pediatric item bank v2.0 Anxiety contains 15 items, the PROMIS pediatric item bank v2.0 Depressive Symptoms contains 14 items, and both short forms 8a contain a subset of eight items. All items use a seven-day recall period and are scored on a five-point Likert scale: 1 (*never*), 2 (*almost never*), 3 (*sometimes*), 4 (*often*), 5 (*almost always*). Level of severity is expressed as theta (θ), and a *T*-score is calculated by the formula (θ × 10) + 50, with higher scores representing higher levels of anxiety or depressive symptoms.

Revised Child Anxiety and Depression Scale short version (RCADS-22) [11, 12, 27]

The RCADS-22 contains 15 items measuring symptoms of anxiety and seven items measuring symptoms of depression in accordance with the DSM-IV [12, 27]. All items are scored on a four-point Likert scale: 0 (*never*), 1 (*sometimes*), 2 (*often*), and 3 (*always*). Total scores are calculated by adding all item scores per subscale, leading to a total score from 0 to 45 on the anxiety subscale and from 0 to 21 on the depression subscale. Higher scores represent a higher level of anxiety or depression. Previous studies have demonstrated strong psychometric properties of the anxiety subscale [12, 27] and the seven items version of the depression subscale [27]. In the present study, the anxiety and depression subscales showed a Cronbach's alpha of 0.87 and 0.84, respectively.

Analyses

We examined whether participants gave identical answers to all RCADS-22 questions and the question with opposite wording, and whether they completed the questionnaire twice for both survey providers. Next, we examined differential item functioning (DIF) for the two samples to assess whether the data could be combined for psychometric analyses.

We performed the following analyses in accordance with the PROMIS analysis plan for psychometric

evaluation of item banks [28]. First, the assumptions of the IRT model were examined: unidimensionality, local independence, and monotonicity. Unidimensionality was examined by confirmatory factor analyses (CFAs) using the R package Lavaan (version 0.6–3) [29]. One-factor model fit was examined using the polychoric correlation matrix with a diagonally weighted least squares estimator. Four fit indices were evaluated: the scaled comparative fit index (*CFI*), the scaled Tucker-Lewis index (*TLI*), the scaled root mean square error of approximation (*RMSEA*), and the standardized root mean square residual (*SRMR*) [30]. Model fit was considered sufficient if the scaled *CFI* and *TLI* > 0.95, the scaled *RMSEA* < 0.06, and *SRMR* < 0.08 [28, 31].

In case of insufficient one-factor model fit, an exploratory bi-factor analysis was examined using the R package psych (version 1.9.12.31) [32]. In a bi-factor model, it is assumed that covariance among item responses can be accounted for by a general factor representing shared variance among all items, and orthogonal group factors representing shared variance over and above the general factor among subsets of items [33, 34]. In case of a strong general factor, an item bank might be considered as unidimensional enough for IRT modeling [33, 34]. Unidimensionality was examined by the Omega-hierarchical (ω_h) and the explained common variance (*ECV*). An $\omega_h > 0.80$ in combination with *ECV* > 0.60 were regarded as indicators of unidimensionalality [35].

Local independence was examined by evaluating residual correlations after controlling for the dominant factor. Residual correlations > 0.20 were considered as indicators of local dependence [28]. Since residual correlations < 0.20 can still lead to model misfit, in addition, we permitted residual correlations with the highest modification indices (*MI*) and examined improvement of model fit. A change of 0.01 for the scaled *CFI* and 0.015 for the scaled *RMSEA* was considered as improved model fit [36].

Monotonicity was examined by a non-parametric IRT model fit with Mokken scaling using the R package Mokken (version 2.8.11) [37]. Model fit was examined by the scalability coefficient *H*. Coefficient $H \geq 0.30$ per item and ≥ 0.50 for the total scale were considered as indicators of an acceptable monotonicity [38].

Second, IRT Graded Response Model (GRM) fit was examined using the R package Mirt (version 1.30) [39]. GRM is an IRT model for ordinal data in which discrimination and threshold parameters are estimated per item using marginal maximum likelihood. The sizes of residuals between observed and expected response frequencies were examined with generalized Orlando and Thissen's $S\text{-}X^2$ statistics for polytomous data; $S\text{-}X^2$ *p* value > 0.001 was considered as an indicator of sufficient item fit [40, 41].

Third, DIF was examined for gender, age group (i.e., aged 8–12 and 13–18), region, ethnicity, and language (i.e., Dutch and English) using the R package Lordif (version 0.3–3) [42]. DIF for language was examined by comparing item responses in our dataset to the dataset PROMIS 1 Pediatric Supplement downloaded at HealthMeasures Dataverse ($N = 1{,}525$, mean age (*SD*) = 12.1 (2.6), girls = 52.1%) [43]. Uniform and non-uniform DIF were examined by ordinal logistic regression models, in which the probability of giving a certain response to an item was modeled as a function of the trait, the group variable, and the interaction of the trait and the group variable. McFadden pseudo $R^2 > 0.02$ was considered as an indication of DIF [42].

In addition to the PROMIS analysis plan, we examined reliability, which is conceptualized as "information" in IRT. Information (*I*) is inversely related to standard errors (*SEs*) and can differ across levels of the measured trait (θ) as indicated by the formula: $SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$. We calculated *SEs* of the short forms and CAT simulations, which were performed using the R package catR (version 3.16) [44]. The CATs started with an item on trait level 0 (which corresponds to a *T*-score of 50) and used the stopping rule of a minimum of four items administered, a maximum of 12 items administered, or an *SE* < 0.316 (which corresponds to a reliability of 0.90).

We calculated *SEs* by using two sets of parameters. The first set of parameters was retrieved from the GRM item fit. We used these Dutch parameters to compare the *SEs* of the short forms and CATs to the *SEs* of the GRM fitted RCADS-22 items per subscale. All *SEs* were plotted on a Dutch metric with a mean *T*-score of 50 and a *SD* of 10 in the Dutch sample. The second set of parameters was the official set of U.S. item parameters in the U.S. calibration sample, obtained from HealthMeasures. We used these U.S. parameters to standardize the *SEs* on the official PROMIS *T*-score metric with a mean *T*-score of 50 and a *SD* of 10 in the U.S. reference sample.

Next, we examined construct validity of the short forms and CATs using SPSS Statistics version 21. We tested predefined hypotheses (following the internationally consensus-based COSMIN checklist [45]). We expected positive correlations ≥ 0.70 between *T*-scores on the Dutch metric and the corresponding RCADS-22 total subscale scores. We expected lower positive correlations between *T*-scores and the non-corresponding RCADS-22 total subscale scores.

Finally, we calculated Dutch reference scores on the universal U.S. PROMIS *T*-score metric per age group (i.e., aged 8–12 and 13–18) and gender in the representative Kantar Public sample and in the total sample. We determined severity cut-offs based on percentiles in the Kantar Public sample [46]: minimal (< 75th percentile), moderate (75–95th percentile), and severe (≥ 95th percentile).

Table 1 Demographic characteristics of the various study samples

	Kantar Public		Total sample $N=2,893$% (% Gold Standard[a])
	Age group 8–12 $N=669$% (% Gold Standard[a])	Age group 13–18 $N=650$% (% Gold Standard[a])	
Gender			
Female	50.1 (48.9)	51.1 (48.9)	52 (48.9)
Age			
8 years	17.8 (19.2)		9.9 (8.5)
9 years	19.7 (19.4)		8.4 (8.7)
10 years	21.4 (20.0)		9.7 (8.7)
11 years	21.2 (20.7)		10.5 (9.0)
12 years	19.9 (20.8)		9.0 (9.3)
13 years		17.4 (16.9)	7.7 (9.3)
14 years		16.9 (17.2)	7.9 (9.4)
15 years		18.5 (16.8)	8.6 (9.6)
16 years		16.8 (16.7)	9.5 (9.3)
17 years		14.2 (16.2)	9.7 (9.3)
18 years		16.3 (16.3)	9.3 (9.0)
Household size			
2 persons or less	4.9 (5.2)	7.6 (7.9)	7.1 (7.2)
3 persons	15.8 (14.3)	18.6 (18.7)	23.7 (16.6)
4 persons	46.3 (43.8)	43.1 (42.4)	43.7 (42.9)
5 persons	23.5 (25.3)	21.4 (20.5)	18.4 (22.8)
6 persons or more	9.4 (10.9)	9.4 (9.5)	7.1 (10.4)
Ethnicity			
Native	78.9 (76.0)	78.0 (76.3)	83.1 (75.3)
First- and second-generation western immigrants[b]	6.1 (6.7)	6.6 (6.8)	5.1 (17.0)
First- and second-generation non-western immigrants[c]	14.9 (17.4)	15.4 (16.9)	11.8 (7.7)
Social class			
Low	8.5 (8.0)	8.0 (8.5)	7.1 (8.2)
Between low and middle	13.8 (13.2)	13.7 (14.3)	13.2 (13.7)
Middle	22.1 (21.2)	24.6 (22.3)	23.7 (23.4)
Between middle and high	23.5 (23.2)	24.8 (24.2)	27 (24.1)
High	32.1 (34.5)	28.9 (30.7)	29 (30.6)
Region			
Three biggest cities[d]	9.9 (10.8)	7.1 (9.5)	9.9 (9.8)
Outskirts of the three biggest cities	4.2 (4.0)	2.5 (3.6)	5.7 (3.7)
West without three biggest cities and outskirts	32.6 (30.1)	31.1 (30.4)	28.6 (30.7)
North	10.3 (10.1)	10.9 (10.3)	12.4 (10.2)
East	22.0 (23.1)	23.8 (22.6)	20.3 (22.8)
South	21.1 (22.0)	24.6 (23.6)	23.2 (22.8)

[a]The Gold Standard is the general Dutch population in 2017

[b]Western = Europe (excluding Turkey), North America, Oceania, Japan, Indonesia (including former Dutch East Indies)

[c]Non-western = Africa, Latin America, Asia (without Japan and Indonesia)

[d]The three biggest cities in the Netherlands are Amsterdam, Rotterdam, The Hague

Results

Sample characteristics

Kantar Public and Panel Inzicht had a response rate of 51 and 39%, respectively. Of 2,933 respondents, 40 were deleted because of conflicting responses on the reverse coded question (Fig. 1). No children participated in both surveys.

The Kantar Public sample was representative per age group 8–12 and 13–18 years: all deviations from the gold standard [26] were < 2.5%, except for native children aged 8–12 (the deviation was 2.9%) (Table 1). No items were flagged for DIF for panel, and the two samples were combined for psychometric analysis ($N = 2{,}893$).

Anxiety

The IRT assumptions were considered to be met. Initially, unidimensionality was partly shown (scaled *CFI* = 0.96; scaled *TLI* = 0.96; scaled *RMSEA* = 0.10; *SRMR* = 0.04; factor loadings varied from 0.71 to 0.91). Since the scaled *RMSEA* was > 0.06, an additional exploratory bifactor analysis was conducted, which yielded high factor loadings on a general factor (0.60 to 0.81). The ω_h was 0.83, and the *ECV* was 0.79, indicating the item bank could be considered as unidimensional enough.

No local dependence was found. Permitting residual correlations between the two items 2230R1r "I got scared really easy" and 227bR1r "I felt afraid" with the highest *MI* (i.e., 482.785) improved model fit (scaled *CFI* = 0.97, scaled *RMSEA* = 0.09). Additionally permitting residual correlations between two different items with the second highest *MI* did not improve model fit anymore.

Monotonicity was considered sufficient; Mokken scalability coefficients of the items ranged from 0.53 to 0.65, and *H* of the full length item bank was 0.61.

All 15 Anxiety items showed sufficient GRM model fit (Table 2). Discrimination parameters ranged from 1.79 to 3.45. Threshold parameters ranged from − 0.10 to 3.52.

No items were flagged for DIF for gender, age group, region, social class, ethnicity, or language.

Figure 2a shows the *SEs* of the full length item bank, short form, CATs, and RCADS-22 anxiety subscale along the *T*-scores scale, calculated with Dutch parameters. The short form showed a *SE* < 3.16 for 51% of the participants, the CATs for 59% of the participants. The CATs used an average of 8.3 items. Item 5044R1r "I felt worried" had the highest discriminating value at $T = 50$ and was therefore administered first in the CATs. The short form and CATs showed a higher reliability over a broader range of *T*-scores than the RCADS-22 anxiety subscale with a smaller (average) number of items.

Figure 2b shows the *SEs* of the full length item bank, short form, and CATs along the official U.S. *T*-score metric. The short form showed a *SE* < 3.16 for 2% of the participants, the CATs for 26% of the participants. Especially participants with *T*-scores < 43 were unreliably estimated (i.e., reliability < 0.80). The CATs used an average of 11.5 items. Item 227bR1r "I felt scared" had the highest discriminating value at $T = 50$ and was therefore administered first in the CATs.

Both hypotheses to examine construct validity were confirmed. Pearson's *r* between the short form and CATs and the RCADS-22 anxiety subscale was 0.75 and 0.74, respectively. The correlations were lower with the RCADS-22 depression subscale: $r = 0.70$ and $r = 0.70$, respectively.

Table 3 shows mean *T*-scores and *SDs* per age group and gender in the representative Kantar Public sample and in the total sample on the official U.S. *T*-score metric. The mean (*SD*) *T*-score of the representative sample was 43.8 (9.9) and varied from 41.1 to 45.5 across subgroups. *T*-scores < 50.77 indicated minimal symptoms, 50.77 ≤ *T*-scores < 61.49 indicated moderate symptoms, and *T*-scores ≥ 61.49 indicated severe symptoms. The mean (*SD*) *T*-score of the total sample was 44.0 (10.5) and varied from 41.5 to 46.2 across subgroups.

Depressive Symptoms

The IRT assumptions were considered to be met. Initially, unidimensionality was partly shown (scaled *CFI* = 0.99; scaled *TLI* = 0.99; scaled *RMSEA* = 0.07; *SRMR* = 0.02; factor loadings varied from 0.72 to 0.94). Since the scaled *RMSEA* was > 0.06, an additional exploratory bifactor analysis was conducted, which yielded high factor loadings on a general factor (0.60–0.90). The ω_h was 0.95, and the *ECV* was 0.93, indicating the item bank could be considered as unidimensional enough.

No local dependence was found. Permitting residual correlations between two items with the highest *MI* did not improve model fit.

Monotonicity was considered sufficient; Mokken scalability coefficients of the items ranged from 0.57 to 0.75, and *H* of the full length item bank was 0.69.

Three out of 14 items did not show sufficient GRM item fit: 2697R1r "I wanted to be by myself", 7010 "I felt sad for no reason", and 9001r "I felt too sad to eat" (Table 2). Item discrimination parameters ranged from 1.82 to 4.86. Threshold parameters ranged from − 0.30 to 3.78.

No items were flagged for DIF for gender, age group, region, social class, and ethnicity, but two items were flagged for uniform DIF for language: 2697R1r "I wanted

Table 2 IRT item characteristics of the PROMIS pediatric Anxiety and Depressive Symptoms item banks in a general Dutch population ($N=2,893$)

Items		Item fit statistics		Discrimination parameters	Difficulty parameters			
		$S\text{-}X^2$	$p\ S\text{-}X^2$	α	β_1	β_2	β_3	β_4
Anxiety								
7005	I felt too nervous to be with a group of children of my age	98.01	0.4522	1.785	0.290	1.368	2.550	3.521
2220R2r	I felt like something awful might happen	71.34	0.3049	3.063	0.575	1.342	2.336	3.426
2230R1r	I got scared really easily	79.38	0.2318	3.413	0.572	1.366	2.194	3.078
227bR1r	I felt scared	88.31	0.0194	3.412	0.457	1.220	2.258	3.433
231R1r	I worried about what could happen to me	78.01	0.2659	3.344	0.444	1.193	2.065	3.060
3021R1r	I was worried I might die	102.10	0.0246	2.737	0.894	1.609	2.501	3.223
3149R1r	I woke up at night scared	118.82	0.0190	2.251	0.768	1.561	2.504	3.358
3150bR2r	I worried when I went to bed at night	103.49	0.1096	2.822	0.615	1.254	1.980	2.818
3459aR1r	I worried when I was away from home	71.45	0.6865	2.877	0.698	1.468	2.300	3.046
3459bR1r	I worried when I was at home	83.70	0.0697	3.448	0.796	1.449	2.259	2.908
3977R1r	I was afraid of going to school	122.59	0.0031	2.911	0.998	1.550	2.260	2.875
5044R1r	I felt worried	100.67	0.1501	2.665	0.071	0.824	1.822	2.659
7006	I worried about what could happen to my parents or caregivers	102.30	0.1966	2.137	0.173	0.936	2.093	3.075
713R1r	I felt nervous	125.00	0.0025	2.150	− 0.103	0.802	1.985	3.326
953R1r	It was hard for me to relax	86.43	0.6442	2.437	0.148	0.947	1.836	2.764
Depressive Symptoms								
2227R1r	I didn't care about anything	99.41	0.4696	1.818	0.305	1.333	2.555	3.781

Table 2 (continued)

Items		Item fit statistics		Discrimination parameters	Difficulty parameters			
		$S\text{-}X^2$	$p\ S\text{-}X^2$	α	β_1	β_2	β_3	β_4
228R1r	I felt sad	113.37	0.0017	2.853	− 0.206	0.660	1,901	3.000
2697R1r	I wanted to be by myself	177.50	0.0000	1.928	− 0.295	0.431	1.853	3.009
3952aR2r	It was hard for me to have fun	73.84	0.2374	3.698	0.339	1.146	2.028	2.739
461R1r	I felt alone	76.63	0.1744	4.133	0.249	0.988	1.776	2.527
488R1r	I could not stop feeling sad	73.56	0.1122	4.861	0.516	1.148	1.889	2.566
5035R1r	I felt like I couldn't do anything right	76.27	0.3738	3.656	0.312	0.978	1.796	2.571
5041R1r	I felt everything in my life went wrong	91.66	0.0927	3.624	0.499	1.160	1.863	2.776
5047R1r	I felt stressed	143.83	0.0022	2.127	− 0.099	0.723	1.786	2.831
679aR2r	Being sad made it hard for me to do things with my friends	99.21	0.0382	3.635	0.542	1.173	1.959	2.555
7010	I felt sad without a reason	123.62	0.0008	3.424	0.590	1.244	2.033	2.740
711R1r	I felt lonely	63.45	0.5661	4.213	0.425	1.092	1.856	2.550
712R1r	I felt unhappy	75.69	0.1505	4.688	0.360	1.012	1.817	2.502
9001r	I felt too sad to eat	175.31	0.0000	2.610	1.047	1.825	2.710	3.380

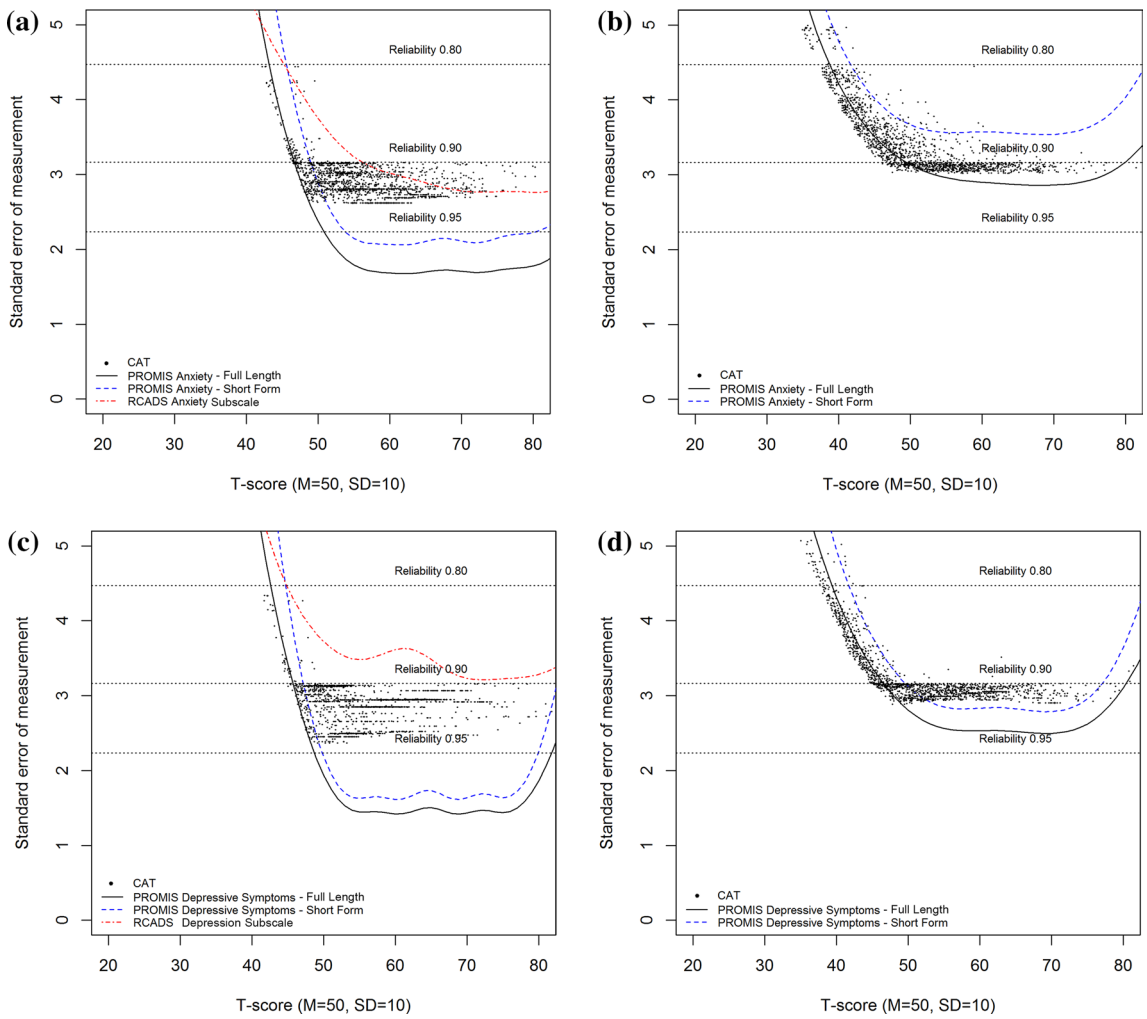

Fig. 2 **a** Standard error of measurement over the range of *T*-scores for the full length Dutch-Flemish PROMIS pediatric item bank v2.0 Anxiety, short form 8a, and CATs, based on Dutch parameters, compared to the RCADS-22 anxiety subscale; **b** Standard error of measurement over the range of *T*-scores for the full length Dutch-Flemish PROMIS pediatric item bank v2.0 Anxiety, short form 8a, and CATs, based on official U.S. parameters; **c** Standard error of measurement over the range of *T*-scores for the Dutch-Flemish PROMIS pediatric item bank v2.0 Depressive Symptoms, short form 8a, and CATs, based on Dutch parameters, compared to the RCADS-22 depression subscale; **d** Standard error of measurement over the range of *T*-scores for the Dutch-Flemish PROMIS pediatric item bank v2.0 Depressive Symptoms, short form 8a, and CATs, based on official U.S. parameters. *CAT* computerized adaptive test; *RCADS* Revised Child Anxiety and Depression Scale; *M* mean; *SD* standard deviation

to be by myself" ($R^2 = 0.030$), and 488R1r "I could not stop feeling sad" ($R^2 = 0.031$).

Figure 2c shows the *SEs* of the full length item bank, short form, CATs, and RCADS-22 depression subscale along the *T*-scores scale, calculated with Dutch parameters. The short form showed a $SE < 3.16$ for 54% of the participants, the CATs for 65% of the participants. The CATs used an average of 6.8 items. Item 461R1r "I felt lonely" had the highest discriminating value at $T = 50$ and was therefore administered first in the CATs. The short form and CATs showed a higher reliability over a broader range of *T*-scores than the RCADS-22 depression subscale.

Figure 2d shows the *SEs* of the full length item bank, short form, and CATs along the official U.S. *T*-score metric. The short form showed a $SE < 3.16$ for 34% of the participants, the CATs for 41% of the participants. Especially participants with *T*-scores < 42 were unreliably estimated (i.e., reliability < 0.80). The CATs used an average of 9.8 items. Item 5035R1r "I felt like I couldn't do anything right" had the highest discriminating value at $T = 50$ and was therefore administered first in the CATs.

Both hypotheses to examine construct validity were confirmed. Pearson's *r* between the short form and CATs and the RCADS-22 depression subscale was 0.78 and 0.76,

Table 3 Mean *T*-scores and standard deviations for each PROMIS pediatric item bank, age group, and gender in a representative[a] general Dutch sample and in the total sample

Age	Gender	Kantar Public sample[a]			Total sample		
		N	Anxiety	Depressive Symptoms	*N*	Anxiety	Depressive Symptoms
			Mean (*SD*)	Mean (*SD*)		Mean (*SD*)	Mean (*SD*)
8–12	Boys	334	44.2 (9.3)	44.6 (9.6)	708	43.7 (9.9)	44.0 (10.2)
	Girls	335	44.2 (9.4)	43.9 (9.9)	662	44.2 (10.3)	44.2 (10.5)
	Total	669	44.2 (9.4)	44.2 (9.7)	1,370	44.0 (10.1)	44.1 (10.3)
13–18	Boys	318	41.1 (9.7)	42.9 (10.5)	681	41.5 (9.9)	43.2 (10.9)
	Girls	332	45.5 (10.6)	47.5 (11.9)	842	46.2 (11.0)	47.9 (12.1)
	Total	650	43.3 (10.4)	45.2 (11.5)	1,523	44.1 (10.8)	45.8 (11.8)
All	Boys	652	42.7 (9.6)	43.8 (10.1)	1,389	42.6 (10.0)	43.6 (10.5)
	Girls	667	44.8 (10.0)	45.7 (11.1)	1,504	45.4 (10.7)	46.3 (11.6)
	Total	1,319	43.8 (9.9)	44.7 (10.6)	2,893	44.0 (10.5)	45.0 (11.2)

[a]Representative sample for each age group 8–12 and 13–18 years old on the variables gender, age, household size, ethnicity, social class, and region (with the exception of native children aged 8–12)

SD standard deviation

respectively. The correlations were lower with the RCADS-22 anxiety subscale: $r = 0.69$ and $r = 0.67$, respectively.

Table 3 shows mean *T*-scores and *SDs* per age group and gender in the representative Kantar Public sample and in the total sample on the official U.S. *T*-score metric. The mean (*SD*) *T*-score of the representative sample was 44.7 (10.6) and varied from 42.9 to 47.5 across subgroups. $T\text{-scores} < 52.78$ indicated minimal symptoms, $52.78 \leq T\text{-scores} < 62.69$ indicated moderate symptoms, and $T\text{-scores} \geq 62.69$ indicated severe symptoms. The mean (*SD*) *T*-score of the total sample was 45.0 (11.2) and varied from 43.2 to 47.9 across subgroups.

Discussion

We evaluated the psychometric properties of the PROMIS pediatric item banks v2.0 Anxiety and Depressive Symptoms, the short forms 8a, and CATs in a general Dutch population. The results support the unidimensionality, local independence, and monotonicity of both item banks and suggest sufficient GRM item fit—except for three Depressive Symptoms items. Both item banks did not show DIF for gender, age group, region, social class, and ethnicity, but two Depressive Symptom items showed DIF for language. With short forms and CATs, reliable scores > 0.80 were obtained for children with moderate and severe levels of anxiety and depression. Construct validity of both short forms and CATs was considered sufficient. Mean *T*-scores for Anxiety and Depressive Symptoms were 43.8 and 44.7 in a representative sample, respectively.

Permission of residual correlation between two Anxiety items with the highest *MI* (i.e., 2230R1r "I got scared really easy" and 227bR1r "I felt afraid") improved model fit, but did not distort parameter estimates. When deleting the item with the lowest discrimination parameter (i.e., 227bR1r "I felt afraid"), discrimination parameters did not change meaningfully (differences ranged from 0.00 to 0.12 and was 0.37 for item 2230R1r "I got scared really easy").

Three Depressive Symptoms items showed poor GRM item fit: 2697R1r "I wanted to be by myself", 7010 "I felt sad for no reason", and 9001r "I felt too sad to eat". These items are not included in the short form but were used in 18% to 63% of the CATs, despite the fact that 7010 "I felt sad for no reason" and 9001r "I felt too sad to eat" had low response curves and therefore a low probability of being selected. A possible explanation is that the Depressive Symptoms item bank consists of only 14 items, and that more informative items measuring similar trait levels are lacking. Also, U.S. discrimination parameters are low, and therefore, almost all items needed to be administered to get a reliable result.

Reliability of the short forms and CATs seemed higher when based on Dutch parameters than when based on U.S. parameters. An explanation might be that more items show DIF for language than presented by Lordif (i.e., 2697R1r "I wanted to be by myself", and 488R1r "I could not stop feeling sad") [47]. Therefore, we additionally examined DIF for language for both item banks using IRT PRO. According to IRT PRO, all Anxiety and Depressive Symptoms items showed DIF for language, except for the Anxiety item 5044R1r "I felt worried". Another explanation might be that calibration samples differed. First, the U.S. calibration sample consisted of a combined general population subset and clinical sample, while the Dutch calibration sample consisted of a general population sample only. This may explain the lower *T*-scores in the Dutch sample (means were 44.0 and 45.0 in the total sample) as compared to the centered average score of 50 in the U.S. calibration sample. The fact that the Dutch calibration sample was a general population sample led to a skewed distribution in scores, which may

have led to inflated discrimination parameters and overestimation of reliability in the Dutch sample [47]. Second, U.S. participants were on average younger than Dutch participants (57.7% versus 47.5% of children aged 8–12 in the U.S. and Dutch sample, respectively). Third, U.S. participants were recruited in person, while Dutch participants were recruited via the internet.

Construct validity of both short forms and CATs was considered sufficient, although differences in correlations between corresponding and non-corresponding constructs were small. These results could be expected given that RCADS short version subscales correlate highly [27, 48].

Strengths of this study are its large sample size and state-of-the art analyses. A limitation is the use of internet survey providers for recruitment of participants, which hampers replication of research procedures; however, it enabled taking a representative sample. Furthermore, the skewed distribution of our data might have caused problems in item parameter estimation. Given the differences between the Dutch and U.S. calibration samples, it is currently not possible to conclude which item parameter set is most appropriate for use in the Dutch population.

The results of this study support the use of both item banks in the Netherlands. Both short forms and CATs showed a reliability > 0.80 for most *T*-scores ≥ 43. The reliability is higher over a broader range in level of anxiety or depression and with fewer items than the reliability of a CTT questionnaire like RCADS-22 [27]. Therefore, both item banks seem useful for assessing and monitoring anxiety and depression in a general population. For now, we recommend the use of U.S. item parameters according to PROMIS convention.

Future research could compare country specific to universal U.S. parameters by collecting additional data in Dutch and U.S. samples using equal inclusion criteria. Furthermore, future research could examine whether more items can be developed, given that both item banks consist of a limited number of items, and three Depressive Symptoms items showed poor GRM item fit.

To conclude, the Dutch-Flemish PROMIS pediatric item banks v2.0 Anxiety and Depressive Symptoms showed sufficient unidimentionality, local independence, monotonicity, and GRM item fit—except for three Depressive Symptom items—in a general Dutch population. DIF for language results were mixed. The short forms 8a and CATs showed sufficient reliability in children with moderate and severe levels of anxiety and depression and sufficient construct validity. More research is needed to examine whether Dutch or U.S. item parameters are optimal for use in the Dutch population.

Acknowledgments We thank B. J. Schalet and A. Kaat for their methodological advice. We are grateful to all participants for spending their time in completing the questionnaire.

Authors contributions All authors contributed to the concept of the study. LK, ML, LH, and CT designed the study. LK, EV, MW, LH, and CT organized the data collection. ML and LK analyzed the data. ML, CT, and LK interpreted the results. LK drafted the manuscript. All authors critically reviewed the manuscript. All authors approved the final version of the manuscript.

Funding The data collection was financially supported by The Netherlands Organization for Health Research and Development (number 729300104). The translation of the PROMIS pediatric item banks was financed by the Dutch-Flemish pediatric PROMIS group.

Data availability The dataset is available on reasonable request from the corresponding author.

Declarations

Conflicts of interest C.B. Terwee is president of the PROMIS Health organization and coordinator of the Dutch-Flemish PROMIS group. C.B. Terwee, L. Haverman, and M.A.J. Luijten are members of the Dutch-Flemish PROMIS group.

Ethics approval The Medical Ethical Committee of the Vrije Universiteit medical center Amsterdam approved the protocol and judged that the Dutch Medical Research Involving Human Subjects Act does not apply to this study. All procedures performed involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent to participate Although the Dutch Medical Research Involving Human Subjects Act does not apply to this study, informed consents were obtained for all participating children.

Consent for publication All authors agree with the publication of this study.

References

1. Polanczyk, G. V., Salum, G. A., Sugaya, L. S., Caye, A., & Rohde, L. A. (2015). Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry, 56*(3), 345–365
2. Erskine, H. E., Moffitt, T. E., Copeland, W. E., Costello, E. J., Ferrari, A. J., Patton, G., et al. (2015). A heavy burden on young minds: The global burden of mental and substance use disorders in children and youth. *Psychological Medicine, 45*(7), 1551–1563
3. Patton, G. C., Coffey, C., Romaniuk, H., Mackinnon, A., Carlin, J. B., Degenhardt, L., et al. (2014). The prognosis of common mental disorders in adolescents: A 14-year prospective cohort study. *The Lancet, 383*(9926), 1404–1411
4. Clayborne, Z. M., Varin, M., & Colman, I. (2019). Systematic review and meta-analysis: Adolescent depression and long-term psychosocial outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry, 58*(1), 72–79
5. Doering, S., Lichtenstein, P., Gillberg, C., Middeldorp, C. M., Bartels, M., et al. (2019). Anxiety at age 15 predicts psychiatric

diagnoses and suicidal ideation in late adolescence and young adulthood: Results from two longitudinal studies. *BMC Psychiatry, 19*, 363
6. Johnson, D., Dupuis, G., Piche, J., Clayborne, Z., & Colman, I. (2017). Adult mental health outcomes of adolescent depression: A systematic review. *Depression and Anxiety, 35*(8), 700–716
7. Siu, A. L. (2016). Screening for depression in children and adolescents: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine, 164*(5), 360–366
8. Stockings, E. A., Degenhardt, L., Dobbins, T., & Lee, Y. Y. (2016). Preventing depression and anxiety in young people: A review of the joint efficacy of universal, selective and indicated prevention. *Psychological Medicine, 46*(1), 11–26
9. Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., et al. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice, 22*(1), 5–19
10. Cohen, J. R., So, F. K., Hankin, B. L., & Lee, B. A. (2019). Youth depression screening with parent and self-reports: Assessing current and prospective depression risk. *Child Psychiatry & Human Development, 50*(4), 647–660
11. Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. *Behaviour Research and Therapy, 38*(8), 835–855
12. Ebesutani, C., Reise, S. P., Chorpita, B. F., Ale, C., Regan, J., Young, J., et al. (2012). The Revised Child Anxiety and Depression Scale-short version: Scale reduction via exploratory bifactor modeling of the broad anxiety factor. *Psychological Assessment, 24*(4), 833–845
13. Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine, 13*, 72
14. McElroy, E., Fearon, P., Belsky, J., Fonagy, P., & Patalay, P. (2018). Networks of depression and anxiety symptoms across development. *Journal of the American Academy of Child & Adolescent Psychiatry, 57*(12), 964–973
15. De Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press.
16. 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
17. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). Initial adult health item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS™) network: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194
18. DeWalt, D. A., Rothrock, N., Yount, S., Stone, A. A., & PROMIS Cooperative Group. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(5 Suppl 1), S12–S21
19. 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
20. 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*(5), 725–735
21. Cella, D., Gershon, R., Lai, J., & 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
22. Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., Morgan DeWitt, E., Lai, J., et al. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research, 19*(4), 595–607
23. Irwin, D. E., Stucky, B. D., Thissen, D., Morgan DeWitt, E., Lai, J., Yeatts, K., et al. (2010). Sampling plan and patient characteristics of the PROMIS pediatrics large-scale survey. *Quality of Life Research, 19*(4), 585–594
24. Alonso, J., Bartlett, S. J., Rose, M., Aaronson, N. K., Chaplin, J. E., Efficace, F., et al. (2013). The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. *Health and Quality of Life Outcomes, 11*(1), 1–5
25. Haverman, L., Grootenhuis, M. A., Raat, H., van Rossum, M. A. J., van Dulmen-den Broeder, E., Hoppenbrouwers, K., et al. (2016). Dutch-Flemish translation of nine pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®. *Quality of Life Research, 25*(3), 761–765
26. MOA. *Gouden standaard: een uniek ijkingsinstrument voor nationale en regionale steekproeven [Gold standard: a unique calibration instrument for national and regional samples]*. https://www.moa.nl/gouden-standaard-expertise-center.html. Accessed 11 September 2020.
27. Klaufus, L., Verlinden, E., van der Wal, M., Kösters, M., Cuijpers, P., & Chinapaw, M. (2020). Psychometric evaluation of two short versions of the Revised Child Anxiety and Depression Scale. *BMC Psychiatry, 20*, 47
28. 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
29. Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software, 48*(2), 1–36
30. Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514
31. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55
32. Revelle, W. (2017). *Psych: procedures for personality and psychological research.* https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research. Accessed 11 September 2020.
33. Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696
34. Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223–237
35. Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5–26
36. Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504
37. Andries van der Ark, L., Koopman, L., Hendrik Straat, J., van den Bergh, D. (2020). *Package 'Mokken'.* http://archive.linux.duke.edu/cran/web/packages/mokken/mokken.pdf. Accessed 11 September 2020.
38. Mokken, R. J. (2011). *A theory and procedure of scale analysis*. Amsterdam: De Gruyter.
39. Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29

40. Kang, T., & Chen, T. T. (2008). Performance of the generalized $S\text{-}X^2$ item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391–406
41. Revicki, D. A., Chen, W., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., et al. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain, 146*(1–2), 158–169
42. 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
43. DeWalt, D. (2016). *PROMIS 1 Pediatric Supplement.* https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IBWSUD. Accessed 11 September 2020.
44. Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software, 48*(8), 1–31
45. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549
46. Carle, A. C., Bevans, K. B., Tucker, C. A., & Forrest, C. B. (2020). Using nationally representative percentiles to interpret PROMIS pediatric measures. *Quality of Life Research, 29*(11), 1–8
47. Kaat, A.J., & Schalet, B.D. (2018). P040 A simulation study of DIF detection procedures, in Proceedings of the 4th Annual PROMIS® Health Organization Conference: Global advances in methodology and clinical science. *Journal of Patient-Reported Outcomes,* 2(Suppl 1): 53.
48. Kösters, M. P., Chinapaw, M. J., Zwaanswijk, M., van der Wal, M. F., & Koot, H. M. (2015). Structure, reliability, and validity of the revised child anxiety and depression scale (RCADS) in a multi-ethnic urban sample of Dutch children. *BMC Psychiatry, 15*, 132