Qual Life Res
DOI 10.1007/s11136-014-0812-7

A new computerized adaptive test advancing the measurement of health-related quality of life (HRQoL) in children: the Kids-CAT

J. Devine · C. Otto · M. Rose · D. Barthel · F. Fischer · H. Mülhan · S. Nolte · S. Schmidt · V. Ottova-Jordan · U. Ravens-Sieberer

Accepted: 18 September 2014

Abstract
Purpose Assessing health-related quality of life (HRQoL) via Computerized Adaptive Tests (CAT) provides greater measurement precision coupled with a lower test burden compared to conventional tests. Currently, there are no European pediatric HRQoL CATs available. This manuscript aims at describing the development of a HRQoL CAT for children and adolescents: the Kids-CAT, which was developed based on the established KIDSCREEN-27 HRQoL domain structure.
Methods The Kids-CAT was developed combining classical test theory and item response theory methods and using large archival data of European KIDSCREEN norm studies ($n = 10{,}577$–$19{,}580$). Methods were applied in line with the US PROMIS project. Item bank development included the investigation of unidimensionality, local independence, exploration of Differential Item Functioning (DIF), evaluation of Item Response Curves (IRCs), estimation and norming of item parameters as well as first CAT simulations.
Results The Kids-CAT was successfully built covering five item banks (with 26–46 items each) to measure physical well-being, psychological well-being, parent relations, social support and peers, and school well-being. The Kids-CAT item banks proved excellent psychometric properties: high content validity, unidimensionality, local independence, low DIF, and model conform IRCs. In CAT simulations, seven items were needed to achieve a measurement precision between .8 and .9 (reliability). It has a child-friendly design, is easy accessible online and gives immediate feedback reports of scores.
Conclusions The Kids-CAT has the potential to advance pediatric HRQoL measurement by making it less burdensome and enhancing the patient–doctor communication.

Electronic supplementary material The online version of this article (doi:10.1007/s11136-014-0812-7) contains supplementary material, which is available to authorized users.

J. Devine (✉) · C. Otto · D. Barthel · V. Ottova-Jordan · U. Ravens-Sieberer
Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics, University Medical Center Hamburg-Eppendorf, Martinistr. 52, W29, 20246 Hamburg, Germany
e-mail: janine.devine@web.de

U. Ravens-Sieberer
e-mail: ravens-sieberer@uke.de
URL: http://www.child-public-health.de

M. Rose · F. Fischer · S. Nolte
Department of Psychosomatic Medicine and Psychotherapy, Charité–University Medicine Berlin, Sauerbruchweg 5, 10117 Berlin, Germany

M. Rose
Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

F. Fischer
Institute for Social Medicine, Epidemiology and Health Economics, Charité University Medicine Berlin, Luisenstr. 57, 10117 Berlin, Germany

H. Mülhan · S. Schmidt
Institute of Psychology, Ernst-Moritz-Arndt Universität Greifswald, Robert-Blum-Str. 13, 17487 Greifswald, Germany

S. Nolte
Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, Burwood, VIC 3125, Australia

Published online: 12 October 2014

Keywords Children · Pediatric · Health-related quality of life · Questionnaire · Item bank · Computerized adaptive test

Introduction

Patient-reported outcomes (PRO) have become an important addition to morbidity indices in pediatric health care. However, pediatric PRO measures are far from being used routinely in clinical practice [1, 2] despite growing consent among clinicians that health-related quality of life (HRQoL) outcomes can aid screening and treatment [3–6]. Because of this special target group, especially when looking at rather young children, the measurement of HRQoL is particularly challenging as children may lose interest in filling out a questionnaire or feel that certain measures are too burdensome. Hence, child-centered measurement may benefit from Computerized Adaptive Tests (CAT) which have proven to be efficient, less burdensome and produce precise and valid scores in adult measurement [7–12]. In particular, they have the potential of easy access online assessments allowing child-friendly test designs and covering the whole spectrum of measurement with a *small* item set that static short forms cannot provide as they are usually either fixed on a limited measurement range or show gaps on the whole measurement spectrum due to a limited number of fixed items.

Due to clear advantages of CATs over static instruments, several research groups started to build pediatric CATs. In the US, pediatric CATs have been developed by Haley et al. [13] and are under development within the Patient Reported Outcomes Measurement Information System (PROMIS®, www.nihpromis.org) Initiative [8, 14–21]). While Haley and colleagues built CATs based on the *established* PEDI [13] to measure physical functioning that have already been evaluated in *longitudinal* studies [14–17, 22–31], the PROMIS initiative constructed *new* items to build 18 CAT item banks for measuring physical, mental and social health in children [18], but they only have been administered to large pediatric *cross-sectional* samples yet [32–37].

Table 1 gives an overview of current pediatric CAT and item banking efforts.

This manuscript aims at presenting the first European effort to develop a pediatric HRQoL CAT in Germany based on large archived national data sets (www.child-public-health.org/deutsch/forschungsinhalte/kids-cat/). The major goals of the Kids-CAT study, funded by the German Federal Ministry of Education and Research (BMBF), are the development of a computerized HRQoL assessment for children and adolescents; the assessment of its score reliability and validity (including responsiveness to change) among children/adolescents with asthma, diabetes and rheumatoid arthritis; and the collection of norm data in healthy children. This manuscript reports on the first goal: the extensive development of the HRQoL CAT called 'Kids-CAT'. It was developed based on the European KIDSCREEN-27 HRQoL domain structure [38–43], aiming at a shorter and more child-friendly, yet equally valid and precise assessment via CAT technology. An additional chronic generic HRQoL item bank complements the Kids-CAT assessing the disease impact of chronic diseases (this item bank development is reported elsewhere). The Kids-CAT will be available online, providing immediate feedback-reporting of the scores to pediatricians.

Methods

The Kids-CAT was developed based on the European KIDSCREEN-27 HRQoL theoretical framework and domain structure [44] with five item banks measuring physical well-being (WB), psychological WB, autonomy and parent relations, social support and peers, and school WB. To do so, we combined classical test theory and item response theory methods following a strategy established by a US research group [45–47], which our German research team adapted and advanced [7, 9, 11, 48–51]. Similar methods have been lately used by the US pediatric PROMIS® project [8, 52, 53].

Samples

Item bank development was based on data from four large European pediatric norm studies: BELLA/KIGGS (t0 or t1: $n = 2{,}863$–$6{,}983$) [40, 54, 55], KIDSCREEN (pilot: $n = 2{,}228$ and norm study: $n = 5{,}108$) [39, 41], HBSC ($n = 5{,}000$) [56–58] and the DISABKIDS ($n = 378$) [43, 59]. For each of the five domains, data from German-speaking countries (Austria, Germany and Switzerland) were combined, resulting in data sets with large sample sizes ranging between 10,577 and 19,580 children/adolescents (for sociodemographics see Table 2). Unlike previous efforts of our research group [9, 51], we chose to merge the study samples *before* the item bank development instead of linking subsamples afterward. This decision was possible, because the available study data consists of large shared "anchor" item subsets (most importantly, constituted by the KIDSCREEN items that allowed for merging).

Construction of the five Kids-CAT item pools

According to the European KIDSCREEN project, pediatric HRQoL can be defined as a "multidimensional construct covering physical, emotional, mental, social and behavioral components of well-being and functioning as perceived by the child" [38–41, 44]. Our research group decided to use the

Table 1 Overview of current pediatric item banks developed for CAT use

No.	Domains	CAT name	Reference	Size of item bank	Sample size	Current status
PROMIS						
1	Anxiety and depression	Pediatric PROMIS® anxiety and depressive symptoms scales	Irwin et al. [93]	18 anxiety and 21 depressive items initially, final item banks: 15 anxiety and 14 depressive items	1,529	Empirical item bank and short form development
2	Anger	PROMIS® Pediatric Anger Scale	Irwin et al. [34]	10 items initially, final item bank: six items	759	Empirical item bank and short form development
3	Stress Response: Somatic and psychological experiences	PROMIS® Pediatric Stress Response item banks	Bevans et al. [95]	2,677 items initially, final item bank: 43 somatic items and 64 psychological items	39	Qualitative item bank development
4	Quality of peer relationships	PROMIS® pediatric peer relationships scale	DeWalt et al. [33]	74 items initially, (53 items: social function, 21 items: sociability), final item bank: 15 items	3,048	Empirical item bank and short form development
5	Six QoL domains (see 6th row)	PROMIS® pediatric item banks	Irwin et al. [97]	293 items initially, final item banks: Physical function: 52 items, Emotional distress: 35 items, Social role relationship: 15 items, Fatigue: 34 items, Pain: 13 items, Asthma: 17 items	4,129	Overview article of six item bank developments (domain-specific articles follow)
6	Physical function: Mobility and upper extremity	PROMIS® pediatric PF item banks	DeWitt et al. [99]	32 mobility and 38 upper extremity items initially, final item banks: 23 mobility and 29 upper extremity items	3,048	Empirical item bank development
7	Mobility	PROMIS® version 1.0 pediatric Mobility CAT	Kratz et al. [100]	Item bank: 23 mobility items, CAT functioning: min. of five items to a max. of 12 items	82 children with cerebral palsy	CAT and short form built and tested for feasibility and validity
8	Fatigue: Tiredness and lack of energy	PROMIS® pediatric fatigue item banks	Lai et al. [36]	39 items initially: 25 tiredness, 14 lack of energy items, final bank: 23 tired and 11 (lack of) energy items	3,048	Empirical item bank and short form development
9	Asthma QoL impact	Pediatric Asthma Impact Scale (PAIS)	Yeatts et al. [37]	34 items initially, final item bank: 17 items	622	Empirical item bank and short form development
Haley research group						
10	Physical function: self-care and mobility	Physical functioning CATs	Haleyet al. [28]	Simulated CATs:–5-item version, –10-item version, –15-item version, –20-item version	373 healthy children, 26 children with Pompe disease	CAT simulated to demonstrate accuracy and the reduction in amount of time
11	Mobility functional skills (of the Pediatric Evaluation of Disability Inventory, PEDI)	Mob-CAT	Haley et al. [25]	Simulated Mob-CAT:–5-item version, - 10-item version, –15-item version, and–59-item full item bank	469 children with disabilities; 412 healthy children	CAT simulated using cross-sectional and longitudinal retrospective data plus small validation study

Table 1 continued

No.	Domains	CAT name	Reference	Size of item bank	Sample size	Current status
12	Self-care and social function	Prototype CAT version of the PEDI	Coster et al. [13]	Self-care item bank: 73 items, social function item bank: 65 items. Simulated Mob-CAT:–5 versus 10 versus 15 item versions	See sample above	CAT simulated using cross-sectional and longitudinal retrospective data plus small validation study
13	Activity in children with cerebral palsy	A new activity item bank	Haley et al. [27]	70 items initially, final item bank: 45 items	308 children with cerebral palsy	CAT simulated with varying stopping rules plus cross-sectional calibration study
14	Physical functioning	CAT for physical functioning of children with cerebral palsy	Tucker et al. [29]	Over 400 items initially, final item banks: Lower extremity skills: 91 items, Upper extremity skills: 53 items, Physical activity: 38 items, Global physical health: 45 items	–	Item bank development

Table 2 Sociodemographics of the data sets used for the Kids-CAT item bank development

Domains	n	Age [mean (SD)]	Male (%)	Germany, Austria, Switzerland (%)	SES [mean (SD); scores 1–5]	Chronic disease or disability (%)
Physical well-being	14,357	13.3 (2.49)	49.5	70, 14, 16	3.9 (0.84)	6.4
Psychological well-being	10,577	12.8 (2.86)	51.8	59, 20, 21	3.5 (1.04)	8.4
Family well-being	19,580	13.2 (2.33)	49.3	78, 10, 11	3.8 (0.85)	4.8
Social well-being	14,366	13.0 (2.38)	48.5	70, 14, 16	3.8 (0.85)	6.6
School well-being	19,300	13.2 (2.31)	49.3	78, 11, 12	3.8 (0.85)	4.9

SES socioeconomic status of the parents rated by the children (1: not good at all, 2: not good, 3: average, 4: good, 5: very good)

existing five well-established pediatric HRQoL domains from the KIDSCREEN project as mentioned before to build a CAT. For definitions of the domains, see Table 3.

For creating the initial item pool, we began with systematic identification and compilation of archived item data of 39 scales like the established HRQoL scales KIDSCREEN [60], KINDL-R [61], CHIP [62], BFW [63], CHQ [64], YQOL [65], plus scales specifically selected for each of the five Kids-CAT domains: for the physical WB item pool: KIGGS [66]; for the psychological WB item pool: the DIKJ [67], CES-DC [68], SCARED [69], CONNERS [70], CBCL [71], CSOC [72], Self Efficacy Scale [73], ACOPE [74] and ECOPE [75]; for the family WB item pool: PBI [76], FKS [77] and HBSC Family Relations Scale [56]; for the social WB item pool: the Oslo Support Scale [78], MOS SSS [79] and HBSC Peer Culture item set [56]; and for the school WB item pool: the HBSC School setting/engagement/achievement, quality of school and school classroom management item sets [56], among others.

Items measuring one of the five HRQoL domains were retrieved in an extensive item selection process scanning all archived studies for eligible items. Items were then sorted to unidimensional item pools by two psychometric experts. The initial item pool started with 495 items. Those items were reviewed in a Delphi process by a team of four psychometric experts. They were asked to review the items to ensure comprehensive coverage of the HRQoL domains as defined in Table 3 and rule out redundant, vague, misclassified, confusing or disease-specific items. Experts rated the appropriateness of each item (yes/no/unsure) independently from each other. Then, items were discussed thoroughly one by one based on the rating results. If the majority of experts (3 out of 4) agreed that respective item covered the content comprehensively, the item was selected for further empirical analysis.

Empirical item analyses and selection

Each of the five item pools underwent careful empirical item analyses and selection covering (a) the investigation of unidimensionality and local independence of all items of each item bank as prerequisite for unidimensional IRT-

Table 3 Description of the Kids-CAT item banks and the content of the excluded items during the whole selection process

Item banks	Definitions	# of initial items before Delphi rating	Reasons for exclusion	# of items after the item selection process
Physical WB	This item bank assesses the child's/adolescent's physical activity, energy, strength, health and fitness as well as the extent to which a child/adolescent feels unwell, complains about poor health or feels sick	72	Items covering specific physical complaints, health care utilization, physical participation, resilience, sleeping problems, appétit or have a time frame >4 weeks	26
Psychological WB	This item bank measures the child's/adolescent's well-being including positive emotions like feeling happy, satisfied with their life, having a purpose in life, self-acceptance and pride—as well as negative emotions like feeling sad, lonely, pressured, worried, insecure, and hopeless	180	Items cover too specific worries, or moods (like anger) related to school/social contexts, ADHS items, items measuring coping behavior or appearance, anhedonia and suicide	46
Family WB	This item bank asks for the interaction between the child/adolescent and parent/carer including whether they feel loved and supported by their family	97	Items covering more social or specific concerns about the family were either excluded or sorted to psyWB and social WB, autonomy items could not be modeled on the same factor	26
Social WB	This item bank measures social relations with friends and peers, the quality and time of interaction between them, and the feeling of being accepted, supported—as well as difficulties finding friends or feeling excluded	75	Items asking about frequency instead of quality of peer relations, items assessing too specific social anxieties or too specific negative interactions like bullying tactics	26
School WB	This item bank assesses the child's/adolescent's perception of his/her cognitive capacity, learning, and concentration and his/her positive and negative feelings about school like feeling happy, satisfied, interested in school versus feeling worried, disappointed, or bored	71	Items asking about school performance (like grades), or too specific school problems	31

based CATs, (b) exploration of Differential Item Functioning (DIF) to rule out severe item bias, (c) evaluation of Item Response Curves (IRCs) to explore whether the response behavior met the IRT function and (d) estimation and norming of item parameters using the Generalized Partial Credit Model (GPCM).

Unidimensionality

Five item banks were constructed using the German CAT algorithm engine already available for an unidimensional CAT [7, 9, 48–51]. Unidimensionality was investigated by Confirmatory Factor Analyses (CFA) based on Full Information Maximum Likelihood (FIML) estimation using the R 2.13, especially package lavaan [80, 81]. To check closely whether this is a solid procedure given the missing data due to block design, we conducted single CFAs of each subset (i.e., data of each original study) and compared the results to the results of the FIML CFAs using the complete merged data set for each dimension. Items were selected using the established criterion of a factor loading higher than .4 [9] referring to Nunnally [82], who observed that factor loadings smaller than .3 should not be taken seriously and that loadings smaller than .4 could easily be over-interpreted. We followed the PROMIS approach to evaluate essential unidimensionality [83, 84].

Local independence is a necessary assumption of the unidimensional IRT model. It means that controlling for trait levels, the response to any item is unrelated to the response to any other item [85]. In other words, there are no other underlying factors explaining the response behavior. To achieve local independence, we examined all residual correlations after fitting a one-factor model. We eliminated one item in each pair of items with a residual correlation of 0.25 or more in line with the criterion applied in earlier studies [9, 50].

Differential Item Functioning (DIF) was analyzed to identify item bias for a wide range of variables like gender, age, education, ethnicity, nationality, socioeconomic and chronic disease status to build non-biased item banks. DIF analyses were conducted using the polytomous logistic regression method [86] on the subsamples of each item pool applying a SAS macro programmed by Bjorner [87]. The sum score of each itempool subset and the above-mentioned variables (gender, age, etc.) were the independent variables, while the item response was the dependent variable in the logistic regression modeling. The criterion of determining DIF was a Nagelkerke's ΔR^2 [88] of $\geq$3 % and $p \geq 0.001$, meaning that the items were discussed and excluded if the independent variable (gender, age, etc.) explained more than 3 % of the variance, and the item was not needed for reasons of content coverage. This criterion has been used as a standard for CAT developments before [9, 49].

Item Response Curves (IRCs) were plotted as a non-parametric method to explore how well the response option curves could be modeled by IRT functions. Criteria to determine the goodness of fit to the IRT modeling were the subsequent order of the response options displayed, the unipolar curve of each IRC, and a mix of steep IRCs (with a high information function of the item on a specific range of the latent trait) and low, widespread IRCs covering the whole latent trait continuum. The IRC modeling was performed using the KernSmoothIRT package provided by the software R [89]. Due to the missings in the block design, the latent trait (x-axis in the Fig. 2) was not modeled by the sum score, but by the rank of the mean sum score of all items that were answered in the specific sample block. Items that did not meet the above requirement, because their IRCs were not in the right order, bipolar or too undiscriminative across the latent trait were deleted.

Item parameters were estimated using the Generalized Partial Credit Model (GPCM) by Muraki [90] by the software Parscale [91]. Like any IRT model, the GPCM models the functioning of item responses by an item response function which describes the probabilistic relation between the responses to an item and the underlying latent trait (called theta), assuming to guide response behavior. We chose to use the GPCM and Parscale because of our previous experience. In GPCM, which is a two-parameter model, the relation is determined by two parameters: the slope parameter (a), giving information about the discriminative ability of an item, and the item threshold parameter (b), indicating the difficulty of an item. The slope parameter is used to estimate the item information function for each item. The parameter estimates are based on a logistic metric. The CAT algorithm used here applies the next item out of the unadministered item bank, which has the highest information function at the current theta estimate. Item fit statistics could not be calculated for the entire sample due to the missings in the block design. Item parameters of the item banks were normed using the representative national KIDSCREEN sample [60], stratifying the sample by age (7–11 vs. 12–17 years old) and gender (male/female) resulting in four subgroups. The stratified norming was done following the recommendation of pediatricians in the tradition of the established KIDSCREEN questionnaires, which provide norm tables for those age and gender groups. Theta scores are natively on a standard normal metric (using a z-score) with a mean of 0 and a standard deviation of 1. For our Kids-CAT, we transformed the scoring to a t-score metric with 50 representing the representative population mean for each of the four subgroups with a standard deviation of 10.

Simulation of the Kids-CAT

First, we simulated new data of 1,000 simulees for each of the five CAT domains. The advantage of simulating new data is that for all items, responses are being simulated to describe the properties of the items of the bank to identify, e.g., ranges of insufficient measurement precision or floor and ceiling effects. We simulated data with a mean of 30 and a SD of 10 to represent a chronically ill population as most items are developed to measure *impairment* of the quality of life in children. Second, we simulated the Kids-CAT using a CAT algorithm programmed by J. Bjorner in SAS. For a description of the CAT process, see [92]. For estimating the scores, the CAT used the Expected A Posteriori method (EAP). The CAT stopped after a maximum of seven items or if a measurement precision of .95 was reached (stopping rule). We simulated CATs for each of the five domains using simulee samples with a mean of 30 and a standard deviation of 10, which were generated at random. We explored the number of items given by the CAT and checked them for content validity and measurement precision across the latent trait [7, 9, 49, 51].

Results

Construction of the five Kids-CAT item pools

During the Delphi process, four psychometric experts were able to select a total of 377 items out of an initial item pool of 495 items from 39 established tools. Forty-four of those items were in the physical WB itempool, 148 items in the psychological WB, 85 items in the autonomy and parental relation, 49 items in the social support and peers, and 51 items in the school WB item pool.

Those items cover the full content range of the five Kids-CAT domains (see Table 3), are child-friendly, comprehensible, and clear in wording, because they stem from scientifically sound established tools. Items that were redundant, vague, misclassified, confusing or disease-specific were excluded. If necessary, item instructions, texts and response options were slightly revised so that they matched to the CAT display: Overall 16 (out of 155 final) CAT items needed slight modifications in the instructions, six items slight modification of the text (e.g., social WB items, which used either children or adolescents in their item wording, were changed to include both words children and adolescents) and three items needed the addition of the response option "I can not answer this question" (which was not scored). The original recall periods (used in the archive studies to build the CAT) were not changed. Items have recall periods between no recall period and 4 weeks recall. We omitted items with a recall period of more than 4 weeks. All recall periods of the original items were kept. The items have 3–6 response options to capture the extent and frequency of the aspect asked for by the item (see Appendix Tables A1–A5 in Supplementary material).

Empirical item analyses and selection

Unidimensionality and local independence

As described in the method section, the block design of the data challenged us in conducting CFA analyses. Figure 1 illustrates that we successfully overcame this challenge by comparing the full information CFA (FIML) approach using the whole data set to CFAs performed in each subset/block. It shows that the two approaches only slightly differ—exemplarily for the physical WB item bank. Thus, we continued conducting CFAs using the FIML approach.

Unidimensionality and local independence were evaluated for all item pools. Only the best items with factor loadings >.4 and residual correlations <.25 were chosen. Initially, we tried to build an item bank to cover both family aspects as well as autonomy and financial resources (like in the KIDSCREEN), but the CFA showed that a unidimensional, solid modeling of family WB autonomy and financial resources need to be excluded.

Then, the CFAs confirmed that all item pools were unidimensional (RMSEA between .03 and .04) and led to 25 items in the physical WB item pool, 81 items in the psychological WB item pool, 39 items in the family WB item bank, 32 items in the social WB item bank and 37 items in the school WB item bank. The item selection is documented in Appendix Tables A1–A5 in Supplementary material.

DIF analyses

Most of the 214 remaining items of the Kids-CAT showed no DIF. Only one item of the physical, nine items of the psychological, one item of the family, three items of the social and no item of the school WB item bank showed DIF using Nagelkerke's $R^2 > 3$ % and/or $p \geq 0.001$.

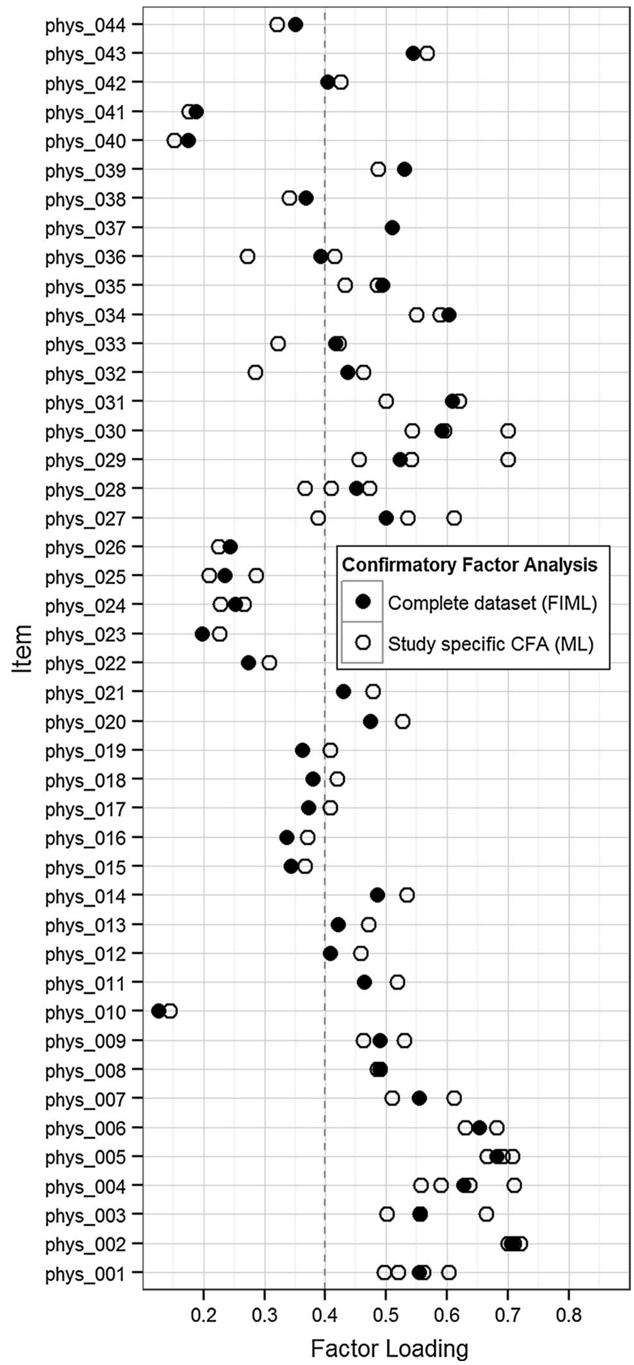

Fig. 1 Comparison of CFA results using the complete data set (FIML) versus the study-specific approach (ML)

Fourteen items showed DIF for age (7–10, 11–13, 14–19 years), gender, ethnicity (16 ethnic categories were differentiated) or social status (measured by the "well-off item" from the child perspective: very well, quite well off, average, not very well, not at all well off). To enhance the item banks, all items displaying gender DIF ("feeling like crying": $R^2 = 6.1$, $p \leq 0.0001$; "feeling sad": $R^2 = 5.2$, $p \leq 0.0001$, "needed to cry": $R^2 = 3.9$, $p \leq 0.0001$, "worried about bad things to happen": $R^2 = 5.2$, $p \leq 0.0001$), ethnicity DIF ("feeling anxious": $R^2 = 3.1$, $p \leq 0.0001$) and social status DIF ("did parents treat you fair?": $R^2 = 13.7$, $p \leq 0.0001$) were excluded. However, four items with age DIF ("I can coordinate my movements": $R^2 = 3.2$, $p \leq 0.0001$; "I was happy": $R^2 = 4.0$, $p \leq 0.0001$; "I felt well": $R^2 = 5.4$, $p \leq 0.0001$; "peers liked me": $R^2 = 5.0$, $p \leq 0.0001$) were kept due to content reasons. To adjust for those differences, we normed the item parameter stratified by age groups. No DIF was found for the variables chronic disease (yes/no) and nationality (German/Austrian/Swiss).

Item Response Curves (IRCs)

Most items showed well-fitting IRCs. Exemplary IRCs of well-fitting items of all item banks are displayed in Fig. 2. The item selection based on the IRCs is thoroughly documented in the Appendix Tables A1–A5 in Supplementary material. In the physical WB item bank, all items met the specified criteria indicating that IRT modeling seemed appropriate. In the psychological WB item bank, most of the items had well-fitting IRCs, and seven items were improved by collapsing their response categories. In the family WB item bank, all items showed good IRCs—except one. In the social support and peers and school WB item banks, most items had well-performing IRCs—except four to five items.

Item parameter estimation and norming

The final Kids-CAT item banks consist of the best performing 26 physical WB, 46 psychological WB, 26 parent relations, 26 social and peers WB and 31 school WB items. Table 3 provides the content of the included and excluded items. It shows that the content coverage of each domain is fully achieved.

The extensive empirical item selection process is documented in the Appendix Tables A1–A5 in Supplementary material. The large-scale norming of the item parameters stratified for boys versus girls and two age groups is displayed for all five item banks in the Appendix Tables B1–B5 in Supplementary material. The tables list all 20 item parameter estimation files, i.e., four item parameter estimations (for boys/girls and two age groups) per item bank (5).

The estimated normed threshold parameters of the physical WB item pool ranged between −6.2 and +1.9, the slope parameters varied between 0.4 and 1.7, the threshold parameters of the psychological WB item pool ranged between −4.4 and +2.5, the slope parameters varied between 0.6 and 1.8, the thresholds of the family WB item pool ranged between −4.3 and +2.2, the slope parameters varied between 0.4 and 2.1, the thresholds of the social WB

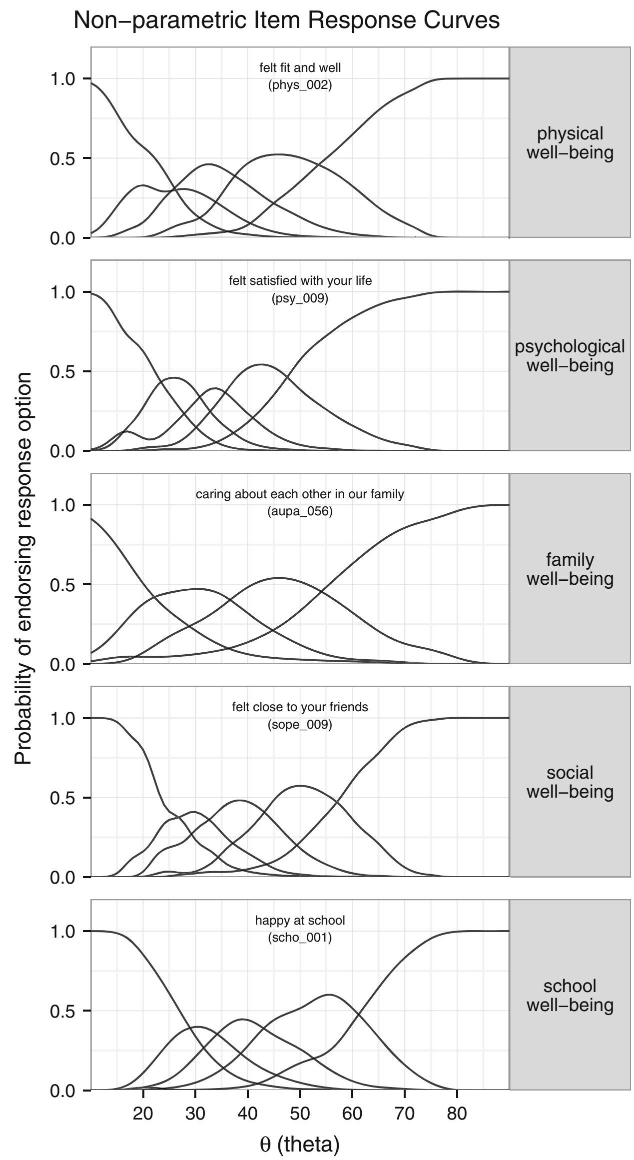

Fig. 2 Examples of items showing regular well-functioning Item Response Curves (IRCs), which were kept during the item selection process

item pool ranged between −3.4 and +1.7, the slope parameters varied between 0.5 and 3.2, the thresholds of the school WB item pool ranged between −4 and +4.1, and the slope parameters varied between 0.5 and 2.6.

Simulations of the Kids-CAT item banks

The five Kids-CAT item banks could be simulated successfully. Figure 3 illustrates the CAT simulations results for the clinical simulee sample (mean = 30, SD = 10): On the *x*-axis, the theta score is displayed on a *t*-score metric, and the *y*-axis shows the measurement precision of each CAT item bank (SE = standard error of measurement). To

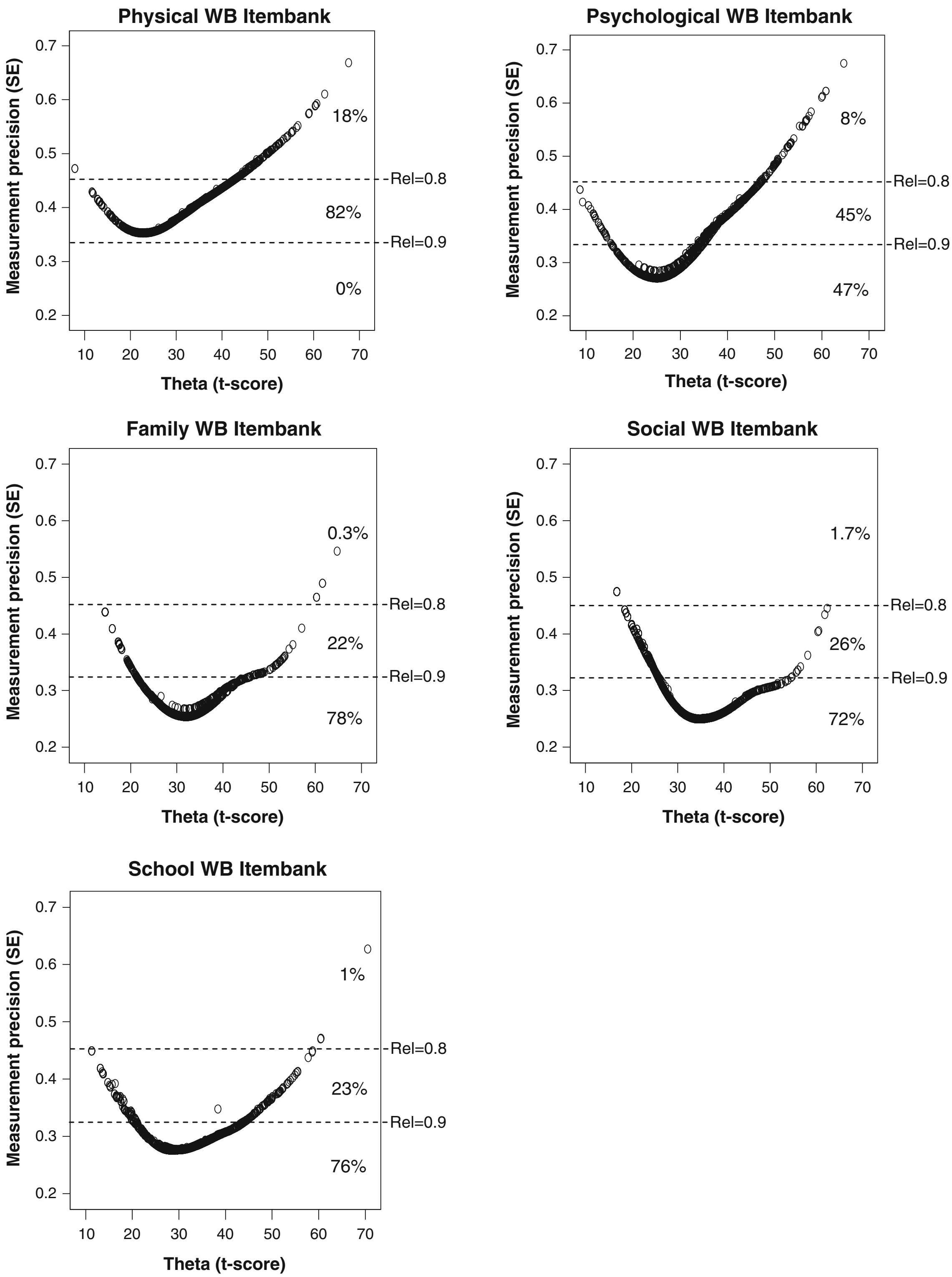

Fig. 3 CAT simulations

ease understanding, the reliability of .8 and .9 are given as horizontal lines in the graphs. The percentages of the 1,000 simulees whose scores had a reliability of <.8, .8 to .9 and >.9 are displayed on the right of each graph.

The Kids-CAT achieved a measurement precision between SE = 0.25 and 0.50 with on average only seven items for the clinical simulee sample. The graphs show that the family, social and school WB item banks have the highest measurement precision (SE = 0.25–0.35) in the range of 30 and 50, i.e., for healthy and impaired children. The psychological and physical WB item banks are slightly less precise (SE = .28–.45) in that range.

When simulating healthy simulee samples (mean = 50, SD = 10), the average number of items needed by the Kids-CAT to achieve a reliability of $\geq$.8 ranges between three (in simulations of the School WB item bank) to seven items (in simulations of the Physical WB item bank).

Overall, the Kids-CAT simulations prove that the Kids-CAT offers a content-valid, precise and low burdensome HRQoL assessment of the child/adolescent.

Discussion

The Kids-CAT is the first *European* CAT measuring generic HRQoL in children and adolescents. The Kids-CAT project combines conceptual and empirical expertise from the KIDSCREEN [38–41, 44], the disease-oriented DISABKIDS [42, 43] and the German adult CAT projects [7, 11, 50, 51].

Major strengths of the Kids-CAT project are that it extensively covers five generic pediatric HRQoL domains, which have already been established as theoretical and empirical framework by the European KIDSCREEN projects, and the item banks are based on 39 scientifically sound established measures and items sets used in various representative archived studies, thus the items can be cross-calibrated to other HRQoL measures. The Kids-CAT was built using large-scale norm data, offers a content-valid, precise, low burden assessment of HRQoL, and measures precise in the range of healthy to impaired children/adolescents, so that it can be applied to healthy and sick children/adolescents. Further, it is easily accessible online, has a child-friendly design, provides immediate easy to interpret score reports and is currently being validated and normed in a healthy representative school sample and in chronically ill children/adolescents.

The Kids-CAT assesses pediatric generic HRQoL CAT covering physical, psychological WB, family WB, social support and peers and school WB as described above (see Table 3).

The *Kids-CAT's Physical WB* item bank (with 26 items) measures the child's/adolescent's physical activity, energy, strength, health and fitness as well as the extent to which a child/adolescent feels unwell, complains about poor health or feels sick.

The *Kids-CAT's Psychological WB* item bank (with 46 items) is a large item bank assessing positive emotions like feeling happy, satisfied with their life, having purpose in life, self-acceptance and pride—as well as negative emotions like feeling sad, lonely, pressured, worried, insecure or hopeless.

The *Kids-CAT's Family WB* item bank (with 26 items) measures the interaction between child/adolescent and parent/carer as well as whether the child/adolescent feels loved and supported by the family. The item bank covers positive family emotions like feeling loved, cared for, supported, and negative family emotions like worrying about other family members or arguing. Initially we tried to build an item bank to cover both family aspects as well as autonomy and financial resources (like in the KIDSCREEN), but the CFA showed that a unidimensional, solid modeling of family WB autonomy and financial resources needs to be excluded.

The *Kids-CAT's Social WB* item bank (with 26 items) assesses the social relations with friends/peers including the quality and time of interaction between them, and the feeling of being accepted, supported—as well as difficulties finding friends or feeling excluded.

The *Kids-CAT's School WB item bank* (with 31 items) measures the child's/adolescent's perception of his/her cognitive capacity including learning, concentration and his/her positive and negative feelings about school like feeling happy, satisfied, interested versus feeling worried, disappointed or bored in school. To date, the school WB item bank is unique in the item banking field, i.e., no US PROMIS pediatric counterpart exists.

A limitation of the Kids-CAT is that it is less precise in the range of the *very* healthy children/adolescents, because the item parameters are most discriminative in the healthy to impaired HRQoL measurement range. Also it needs to be added that the CAT simulations were performed on simulated data, which were simulated assuming the validity of the model, thus the CAT simulation results are likely more favorable than they would be in real CAT-applications.

Comparing the European Kids-CAT to the US pediatric counterparts: It is similar to the PEDI-CAT project in that the Kids-CAT was built using established tools. It is different in that the PEDI-CATs measure physical mobility/activity and self-care by *proxy-report*, while the Kids-CAT offers *self-report* assessment of children and adolescents. And while most of the PROMIS item banks are more *symptom*-oriented measuring physical functioning of the upper extremity/mobility [16, 17, 25–28, 35], emotional distress (depression/anxiety [93], anger [94], stress [95]), fatigue, pain (quality/interference) and asthma impact [34, 93, 94, 96, 97], the Kids-CAT covers pediatric general HRQoL more broadly—targeting healthy and sick/impaired children/adolescents like the pediatric PROMIS efforts on subjective well-being (SWB [32, 98]). However, the Kids-CAT is based on a more established theoretical and empirical background: the domain structure follows the large European KIDSCREEN project, the item banks are drawn from established instruments and built using existing large-scale German norm data from the KIDSCREEN, BELLA/KIGGS, HBSC and DISABKIDS studies, while the PROMIS projects created items from scratch (with no intial database). Hence, the Kids-CAT item banks can be linked and equated to previous and future studies, i.e., international comparisons are facilitated.

To summarize: this manuscript illustrated the successful quantitative development of the Kids-CAT using large-scale European norm data sets from German-speaking countries and a solid IRT-based methodological approach. The Kids-CAT covers the most important domains of pediatric generic HRQoL in line with the KIDSCREEN and the DISABKIDS. This manuscript is followed by a future manuscript that will describe the qualitative Kids-CAT item evaluations and CAT programming. Currently, the Kids-CAT is being administered to a norm sample of 1,200 German school children and to a clinical sample of 300 children with chronic diseases (asthma, diabetes, rheumatoid arthritis) across two German pediatric centers (Kiel, Lübeck). Those ongoing studies aim at evaluating the reliability, validity and responsiveness to change of the Kids-CAT. In future studies, the Kids-CAT will be normed and a User's guide will be published to facilitate score interpretations. The User's guide will include a CD and online access, so that pediatricians can easily administer the Kids-CAT and implement it into routine pediatric care.

Conclusions

The five Kids-CAT item banks (with 26–46 items per bank) show good psychometric properties, that is, high content validity, sufficient unidimensionality and local independence, no significant DIF and regular IRCs, allowing for item parameter estimation. First Kids-CAT simulation results are promising: seven items are displayed with a reliability of .8 to .9. The Kids-CAT has the potential to advance pediatric HRQoL measurement by easy administration, scoring and immediate feedback-reporting.

Acknowledgments This work was funded by the German Federal Ministry of Education and Research (BMBF, Grant 0010-01GY1111, PI: Prof. Dr. Ulrike Ravens-Sieberer, MPH, University Medical Center Hamburg-Eppendorf). We would like to thank our advisory board members (Prof. Dr. Christopher Forrest, Prof. Dr. Lena Lämmle, Prof. Dr. Markus Wirtz, and Prof. Dr. Monika Bullinger) for the helpful advice and support. We also thank all children and parents, who participated in the archived studies, which were used for building the Kids-CAT.

References

1. Clarke, S. A., & Eiser, C. (2004). The measurement of health-related quality of life (QOL) in paediatric clinical trials: A systematic review. *Health and Quality of Life Outcomes, 2*, 66.
2. Solans, M., Pane, S., Estrada, M. D., Serra-Sutton, V., Berra, S., Herdman, M., et al. (2008). Health-related quality of life measurement in children and adolescents: A systematic review of generic and disease-specific instruments. *Value in Health, 11*, 742–764.
3. Detmar, S. B., Muller, M. J., Schornagel, J. H., Wever, L. D., & Aaronson, N. K. (2002). Health-related quality-of-life assessments and patient-physician communication: A randomized controlled trial. *JAMA, 288*, 3027–3034.
4. Engelen, V., Detmar, S., Koopman, H., Maurice-Stam, H., Caron, H., Hoogerbrugge, P., et al. (2012). Reporting health-related quality of life scores to physicians during routine follow-up visits of pediatric oncology patients: Is it effective? Pediatr. *Blood Cancer, 58*, 766–774.
5. Gutteling, J. J., Darlington, A. S., Janssen, H. L., Duivenvoorden, H. J., Busschbach, J. J., & de Man, R. A. (2008). Effectiveness of health-related quality-of-life measurement in clinical practice: A prospective, randomized controlled trial in patients with chronic liver disease and their physicians. *Quality of Life Research, 17*, 195–205.
6. de la Osa, N., Ezpeleta, L., Granero, R., & Domenech, J. M. (2009). Brief mental health screening questionnaire for children and adolescents in primary care settings. *International Journal of Adolescent Medicine and Health, 21*, 91–100.
7. Becker, J., Fliege, H., Kocalevent, R. D., Bjorner, J. B., Rose, M., Walter, O. B., et al. (2008). Functioning and validity of a computerized adaptive test to measure anxiety (A-CAT). *Depression and Anxiety, 25*, E182–E194.
8. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*, 1179–1194.
9. Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*, 2277–2291.
10. Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., et al. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry, 69*, 1104–1112.
11. Rose, M., Bjorner, J. B., Fischer, F., Anatchkova, M., Gandek, B., Klapp, B. F., et al. (2012). Computerized adaptive testing–ready for ambulatory monitoring? *Psychosomatic Medicine, 74*, 338–348.
12. Turner-Bowker, D. M., Saris-Baglama, R. N., Smith, K. J., DeRosa, M. A., Paulsen, C. A., & Hogue, S. J. (2011). Heuristic evaluation and usability testing of a computerized patient-reported outcomes survey for headache sufferers. *Telemedicine Journal and E-Health, 17*, 40–45.
13. Coster, W. J., Haley, S. M., Ni, P., Dumas, H. M., & Fragala-Pinkham, M. A. (2008). Assessing self-care and social function using a computer adaptive testing version of the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation, 89*, 622–629.
14. Dumas, H. M., Fragala-Pinkham, M. A., Haley, S. M., Ni, P., Coster, W., Kramer, J. M., et al. (2012). Computer adaptive test performance in children with and without disabilities: Prospective field study of the PEDI-CAT. *Disability and Rehabilitation, 34*, 393–401.

15. Dumas, H. M., & Fragala-Pinkham, M. A. (2012). Concurrent validity and reliability of the pediatric evaluation of disability inventory-computer adaptive test mobility domain. *Pediatric Physical Therapy, 24*, 171–176.
16. Haley, S. M., Raczek, A. E., Coster, W. J., Dumas, H. M., & Fragala-Pinkham, M. A. (2005). Assessing mobility in children using a computer adaptive testing version of the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation, 86*, 932–939.
17. Haley, S. M., Chafetz, R. S., Tian, F., Montpetit, K., Watson, K., Gorton, G., et al. (2010). Validity and reliability of physical functioning computer-adaptive tests for children with cerebral palsy. *Journal of Pediatric Orthopedics, 30*, 71–75.
18. Forrest, C. B., Bevans, K. B., Tucker, C., et al. (2012). The patient reported outcome measurement information system (PROMIS®) for children and youth: Application to pediatric psychology. *Journal of Pediatric Psychology, 37*, 614–621.
19. Haley, S. M., Ni, P., Ludlow, L. H., & Fragala-Pinkham, M. A. (2006). Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation, 87*, 1223–1229.
20. Haley, S. M., Coster, W. J., Dumas, H. M., Fragala-Pinkham, M. A., Kramer, J., Ni, P., et al. (2011). Accuracy and precision of the pediatric evaluation of disability inventory computer-adaptive tests (PEDI-CAT). *Developmental Medicine and Child Neurology, 53*, 1100–1106.
21. 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*, S3–S11.
22. Dumas, H., Fragala-Pinkham, M., Haley, S., Coster, W., Kramer, J., Kao, Y. C., et al. (2010). Item bank development for a revised pediatric evaluation of disability inventory (PEDI). *Physical & Occupational Therapy in Pediatrics, 30*, 168–184.
23. Dumas, H. M., Fragala-Pinkham, M. A., & Haley, S. M. (2010). Development of a postacute hospital item bank for the new pediatric evaluation of disability inventory-computer adaptive test. *International Journal of Rehabilitation Research, 33*, 332–338.
24. Dumas, H. M., Rosen, E. L., Haley, S. M., Fragala-Pinkham, M. A., Ni, P., & O'Brien, J. E. (2010). Measuring physical function in children with airway support: A pilot study using computer adaptive testing. *Developmental Neurorehabilitation, 13*, 95–102.
25. Haley, S. M., Ni, P., Fragala-Pinkham, M. A., Skrinar, A. M., & Corzo, D. (2005). A computer adaptive testing approach for assessing physical functioning in children and adolescents. *Developmental Medicine and Child Neurology, 47*, 113–120.
26. Haley, S. M., Fragala-Pinkham, M., & Ni, P. (2006). Sensitivity of a computer adaptive assessment for measuring functional mobility changes in children enrolled in a community fitness programme. *Clinical Rehabilitation, 20*, 616–622.
27. Haley, S. M., Fragala-Pinkham, M. A., Dumas, H. M., Ni, P., Gorton, G. E., Watson, K., et al. (2009). Evaluation of an item bank for a computerized adaptive test of activity in children with cerebral palsy. *Physical Therapy, 89*, 589–600.
28. Haley, S. M., Ni, P., Dumas, H. M., Fragala-Pinkham, M. A., Hambleton, R. K., Montpetit, K., et al. (2009). Measuring global physical health in children with cerebral palsy: Illustration of a multidimensional bi-factor model and computerized adaptive testing. *Quality of Life Research, 18*, 359–370.
29. Tucker, C. A., Haley, S. M., Dumas, H. M., Fragala-Pinkham, M. A., Watson, K., Gorton, G. E., et al. (2008). Physical function for children and youth with cerebral palsy: Item bank development for computer adaptive testing. *Journal of Pediatric Rehabilitation Medicine, 1*, 245–253.
30. Tucker, C. A., Gorton, G. E., Watson, K., Fragala-Pinkham, M. A., Dumas, H. M., Montpetit, K., et al. (2009). Development of a parent-report computer-adaptive test to assess physical functioning in children with cerebral palsy I: Lower-extremity and mobility skills. *Developmental Medicine and Child Neurology, 51*, 717–724.
31. Tucker, C. A., Montpetit, K., Bilodeau, N., Dumas, H. M., Fragala-Pinkham, M. A., Watson, K., et al. (2009). Development of a parent-report computer-adaptive test to assess physical functioning in children with cerebral palsy II: Upper-extremity skills. *Developmental Medicine and Child Neurology, 51*, 725–731.
32. Bevans, K. B., Riley, A. W., & Forrest, C. B. (2010). Development of the healthy pathways child-report scales. *Quality of Life Research, 19*, 1195–1214.
33. DeWalt, D. A., Thissen, D., Stucky, B. D., Langer, M. M., Morgan, D. E., Irwin, D. E., et al. (2013). PROMIS pediatric peer relationships scale: Development of a peer relationships item bank as part of social health measurement. *Health Psychology, 32*, 1093–1103.
34. Irwin, D. E., Gross, H. E., Stucky, B. D., Thissen, D., Dewitt, E. M., Lai, J. S., et al. (2012). Development of six PROMIS pediatrics proxy-report item banks. *Health and Quality of Life Outcomes, 10*, 22.
35. Kerfeld, C. I., Dudgeon, B. J., Engel, J. M., & Kartin, D. (2013). Development of items that assess physical function in children who use wheelchairs. *Pediatric Physical Therapy, 25*, 158–166.
36. Lai, J. S., Stucky, B. D., Thissen, D., Varni, J. W., Dewitt, E. M., Irwin, D. E., et al. (2013). Development and psychometric properties of the PROMIS (R) pediatric fatigue item banks. *Quality of Life Research, 22*, 2417–2427.
37. Yeatts, K. B., Stucky, B., Thissen, D., Irwin, D., 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.
38. Ravens-Sieberer, U., Schmidt, S., Gosch, A., Erhart, M., Petersen, C., & Bullinger, M. (2007). Measuring subjective health in children and adolescents: Results of the European KIDSCREEN/DISABKIDS Project. Psychosoc. Med., 4, Doc08.
39. Ravens-Sieberer, U., Auquier, P., Erhart, M., Gosch, A., Rajmil, L., Bruil, J., et al. (2007). The KIDSCREEN-27 quality of life measure for children and adolescents: Psychometric results from a cross-cultural survey in 13 European countries. *Quality of Life Research, 16*, 1347–1356.
40. Ravens-Sieberer, U., Erhart, M., Gosch, A., & Wille, N. (2008). Mental health of children and adolescents in 12 European countries-results from the European KIDSCREEN study. *Clinical Psychology and Psychotherapy, 15*, 154–163.
41. Ravens-Sieberer, U., Herdman, M., Devine, J., Otto, C., Bullinger, M., Rose, M., et al. (2013). The European KIDSCREEN approach to measure quality of life and well-being in children: Development, current application, and future advances. *Quality of Life Research, 23*, 791–803.
42. Schmidt, S., Debensason, D., Muhlan, H., Petersen, C., Power, M., Simeoni, M. C., et al. (2006). The DISABKIDS generic quality of life instrument showed cross-cultural validity. *Journal of Clinical Epidemiology, 59*, 587–598.
43. Schmidt, S., Thyen, U., Chaplin, J., Mueller-Godeffroy, E., & Bullinger, M. (2008). Healthcare needs and healthcare satisfaction from the perspective of parents of children with chronic conditions: The DISABKIDS approach towards instrument development. *Child Care, Health and Development, 34*, 355–366.

44. Ravens-Sieberer, U., Gosch, A., Rajmil, L., Erhart, M., Bruil, J., Duer, W., et al. (2005). KIDSCREEN-52 quality-of-life measure for children and adolescents. *Expert Review of Pharmacoeconomics & Outcomes Research, 5*, 353–364.
45. Ware, J. E., Jr., Bjorner, J. B., & Kosinski, M. (1999). Dynamic health assessment: The search for more practical and more precise outcome measures. *Quality of Life Newsletter*, 11–13.
46. Ware, J. E., Jr., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38*, II73–II82.
47. Bjorner, J. B., Kosinski, M., & Ware, J. E, Jr. (2003). Using item response theory to calibrate the Headache Impact Test (HIT™) to the metric of traditional headache scales. *Quality of Life Research, 12*, 981–1002.
48. Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B., & Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research, 18*, 23–36.
49. Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., et al. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology, 62*(278–87), 287.
50. Walter, O. B., Becker, J., Fliege, H., Bjorner, J., Kosinski, M., Walter, M., et al. (2005). Entwicklungsschritte fuer einen computeradaptiven Test zur Erfassug von Angst (A-CAT). [Developmental steps for a computer-adapted test for anxiety]. *Diagnostica, 51*, 88–100.
51. Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Quality of Life Research, 16*(Suppl 1), 143–155.
52. Forrest, C. B. (2013). Advancing pediatric patient-reported outcome assessment. *Value Health, 16*, 907–908.
53. Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology, 23*, S53–S57.
54. Kurth, B. M., Kamtsiuris, P., Holling, H., Schlaud, M., Dolle, R., Ellert, U., et al. (2008). The challenge of comprehensively mapping children's health in a nation-wide health survey: Design of the German KiGGS-study. *BMC Public Health, 8*, 196.
55. Ravens-Sieberer, U., Erhart, M., Wille, N., & Bullinger, M. (2008). Health-related quality of life in children and adolescents in Germany: Results of the BELLA study. *European Child and Adolescent Psychiatry, 17*(Suppl 1), 148–156.
56. Currie, C., Nic, G. S., & Godeau, E. (2009). The Health Behaviour in School-aged Children: WHO Collaborative Cross-National (HBSC) study: Origins, concept, history and development 1982–2008. *International Journal of Public Health, 54*(Suppl 2), 131–139.
57. Ottova, V., Hillebrandt, D., & Ravens-Sieberer, U. (2012). Trends in subjective health and well-being of children and adolescents in Germany: Results of the Health Behaviour in School-aged Children (HBSC) Study 2002 to 2010. *Gesundheitswesen, 74*(Suppl), S15–S24.
58. Ravens-Sieberer, U. (2009). The contribution of HBSC to international child health research: A milestone in child public health. *International Journal of Public Health, 54*(Suppl 2), 121–122.
59. Baars, R. M., Atherton, C. I., Koopman, H. M., Bullinger, M., & Power, M. (2005). The European DISABKIDS project: Development of seven condition-specific modules to measure health related quality of life in children and adolescents. *Health and Quality of Life Outcomes, 3*, 70.
60. The KIDSCREEN GROUP EUROPE. (2006). *The KIDSCREEN questionnaires - Quality of life questionnaires for children and adolescents*. Lengerich: Pabst.
61. Ravens-Sieberer, U., & Bullinger, M. (1998). Assessing health-related quality of life in chronically ill children with the German KINDL: First psychometric and content analytical results. *Quality of Life Research, 7*, 399–407.
62. Starfield, B., Riley, A. W., Forrest, C. B., Green, B. F., Robertson, J. A., & Rajmil, L. (2007). Child Health and Illness Profile (CHIP). A comprehensive assessment of health and functioning of children and adolescents. Johns Hopkins Bloomberg School of Public Health.
63. Grob, A., Lüthi, R., Kaiser, F. G., Flammer, A., Mackinnon, A., & Wearing, A. J. (1991). Berner Fragebogen zum Wohlbefinden Jugendlicher (BFW) (Bernese questionnaire of subjective well-being). *Diagnostica, 37*, 66–75.
64. The Child Health Questionnaire (CHQ). (2000). *Scoring and Interpretation Manual*. Boston: HealthActCHQ Inc.
65. Topolski, T. D., Edwards, T. C., & Patrick, D. L. (2002). *User's manual and interpretation guide for the youth quality of life (YQOL) instruments*. University of Washington, Department of Health Services, Seattle, WA.
66. Kurt, B. M. (2005). KIGGS. The German health survey for children and adolescents. Robert Koch Institute, Head Department of Epidemiology and Health Reporting.
67. Stiensmeier-Pelster, J., Schürmann, M., & Duda, K. (1989). *Depressions-Inventar für Kinder und Jugendliche (DIKJ)*. Göttingen: Hogrefe.
68. Faulstich, M. E., Carey, M. P., Ruggiero, L., Enyart, P., & Gresham, F. (1986). Assessment of depression in childhood and adolescence: An evaluation of the Center for Epidemiological Studies Depression Scale for Children (CES-DC). *American Journal of Psychiatry, 143*, 1024–1027.
69. Birmaher, B., Brent, D. A., Chiappetta, L., Bridge, J., Monga, S., & Baugher, M. (1999). Psychometric properties of the screen for childhood anxiety related emotional disorders (SCARED): A replication study. *Journal of the American Academy of Child and Adolescent Psychiatry, 38*(10), 1230–1236.
70. Conners, C. K. (2008). *Conners scale* (3rd ed.). North Tonawanda, NY: Multi-Health Systems Inc.
71. Achenbach, T. M., & Rescorla, L. (2001). *ASEBA school-age forms and profiles*. Burlington: Aseba.
72. Margalit, M. (1995). *CSOC: Children sense of coherence manual*. Tel Aviv: Tel Aviv University Press.
73. Jerusalem, M., & Schwarzer, R. (1999). Allgemeine Selbstwirksamkeit. In R. Schwarzer & M. Jerusalem (Eds.), *Skalen zur Erfassung von Lehrer-und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin: Free University of Berlin.
74. Jerusalem, M. & Mittag, W. (1999). Problemorientiertes, aktives Coping (ACOPE). In R. Schwarzer & M. Jerusalem (Eds.), Skalen zur Erfassung von Lehrer-und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen. Berlin: Free University of Berlin.
75. Bäßler, J., & Schwarzer, R. (1999). Emotionsorientiertes, vermeidendes Coping (ECOPE). In R. Schwarzer & M. Jerusalem (Eds.), Skalen zur Erfassung von Lehrer-und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen. Berlin: Free University of Berlin.
76. Parker, G., Tupling, H., & Brown, L. B. (1979). Parental bonding instrument (PBI). *British Journal of Medical Psychology, 52*, 1–10.

77. Schneewind, K. A. (2014). Die Familienklimaskalen (FKS). In M. Cierpa (Ed.), *Familiendiagnostik* (pp. 232–255). Berlin: Springer.
78. Meltzer, H. (2003). Development of a common instrument for mental health. In A. Nosikov & C. Gudex (Eds.), *EUROHIS: Developing common instruments for health surveys*. Amsterdam: IOS Press.
79. Sherbourne, C. D., & Steward, A. L. (1991). The MOS social support survey. *Social Science and Medicine, 32*, 705–714.
80. R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.
81. Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36.
82. Nunnally, J. (1978). *Psychometric Theory* (2nd ed.). New York: MacGraw-Hill.
83. Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*, 447–460.
84. 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*, S22–S31.
85. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum Associates.
86. Swaminathan, H., & Rogers, J. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.
87. Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology, 51*, 1189–1202.
88. Nagelkerke, N. J. D. (1991). Miscellanea. A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.
89. Mazza, A., Punzo, A., & McGuire, B. (2012). KernSmoothIRT: AN R package for kernel smoothing in item response theory. Cornell University Library. Ref Type: Electronic Citation.
90. Muraki, E. (1997). A generalized partial credit model. In W. J. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). Berlin: Springer.
91. Muraki, E., & Bock, R. D. (1999). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago: Scientific Software Int.
92. Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., et al. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.
93. Irwin, D. E., Stucky, B., 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.
94. Irwin, D. E., Stucky, B. D., Langer, M. M., Thissen, D., Dewitt, E. M., Lai, J. S., et al. (2011). PROMIS® pediatric anger scale: An item response theory analysis. *Quality of Life Research, 21*(4), 697–706.
95. Bevans, K. B., Gardner, W., Pajer, K., Riley, A. W., & Forrest, C. B. (2013). Qualitative development of the PROMIS(R) pediatric stress response item banks. *Journal of Pediatric Psychology, 38*, 173–191.
96. 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.
97. 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.
98. Ravens-Sieberer, U., Devine, J., Bevans, K., Riley, A. W., Moon, J., Salsman, J. M., et al. (2014). Subjective well-being measures for children were developed within the PROMIS project: Presentation of first results. *Journal of Clinical Epidemiology, 67*, 207–218.
99. DeWitt, E. M., Stucky, B. D., Thissen, D., Irwin, D. E., Langer, M., Varni, J. W., ... & DeWalt, D. A. (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*(7), 794–804.
100. Kratz, A. L., Slavin, M. D., Mulcahey, M. J., Jette, A. M., Tulsky, D. S., & Haley, S. M. (2013). An examination of the PROMIS® pediatric instruments to assess mobility in children with cerebral palsy. *Quality of Life Research, 22*(10), 2865–2876.