

# The validation of a computer-adaptive test (CAT) for assessing health-related quality of life in children and adolescents in a clinical sample: study design, methods and first results of the Kids-CAT study

D. Barthel<sup>1</sup> · C. Otto<sup>1</sup> · S. Nolte<sup>2,3</sup> · A.-K. Meyrose<sup>1</sup> · F. Fischer<sup>2,4</sup> ·  
J. Devine<sup>1</sup> · O. Walter<sup>2</sup> · A. Mierke<sup>2</sup> · K. I. Fischer<sup>2</sup> · U. Thyen<sup>5</sup> · M. Klein<sup>6</sup> ·  
T. Anker mann<sup>6</sup> · M. Rose<sup>2,7</sup> · U. Ravens-Sieberer<sup>1</sup>

Accepted: 12 October 2016 / Published online: 9 November 2016  
© Springer International Publishing Switzerland 2016

## Abstract

**Purpose** Recently, we developed a computer-adaptive test (CAT) for assessing health-related quality of life (HRQoL) in children and adolescents: the Kids-CAT. It measures five generic HRQoL dimensions. The aims of this article were (1) to present the study design and (2) to investigate its psychometric properties in a clinical setting.

**Methods** The Kids-CAT study is a longitudinal prospective study with eight measurements over one year at two University Medical Centers in Germany. For validating the Kids-CAT, 270 consecutive 7- to 17-year-old patients with asthma ( $n = 52$ ), diabetes ( $n = 182$ ) or juvenile arthritis ( $n = 36$ ) answered well-established HRQoL instruments (Pediatric Quality of Life Inventory<sup>TM</sup> (PedsQL), KIDSCREEN-27) and scales measuring related constructs

(e.g., social support, self-efficacy). Measurement precision, test–retest reliability, convergent and discriminant validity were investigated.

**Results** The mean standard error of measurement ranged between .38 and .49 for the five dimensions, which equals a reliability between .86 and .76, respectively. The Kids-CAT measured most reliably in the lower HRQoL range. Convergent validity was supported by moderate to high correlations of the Kids-CAT dimensions with corresponding PedsQL dimensions ranging between .52 and .72. A lower correlation was found between the social dimensions of both instruments. Discriminant validity was confirmed by lower correlations with non-corresponding subscales of the PedsQL.

**Conclusions** The Kids-CAT measures pediatric HRQoL reliably, particularly in lower areas of HRQoL. Its test–retest reliability should be re-investigated in future studies. The validity of the instrument was demonstrated. Overall,

**Electronic supplementary material** The online version of this article (doi:10.1007/s11136-016-1437-9) contains supplementary material, which is available to authorized users.

✉ U. Ravens-Sieberer  
ravens-sieberer@uke.de

<sup>1</sup> Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics, Research Unit Child Public Health, Center for Psychosocial Medicine, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany

<sup>2</sup> Department of Psychosomatic Medicine, Center of Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

<sup>3</sup> Public Health Innovation, Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, 221 Burwood Highway, Melbourne, VIC 3125, Australia

<sup>4</sup> Institute for Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Luisenstraße 57, 10117 Berlin, Germany

<sup>5</sup> Hospital for Pediatrics and Adolescent Medicine, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

<sup>6</sup> Department of Pediatrics, University Medical Center Schleswig-Holstein, Campus Kiel, Arnold-Heller-Straße 3, House 9, 24105 Kiel, Germany

<sup>7</sup> Department of Quantitative Health Sciences, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605, USA

results suggest that the Kids-CAT is a promising candidate for detecting psychosocial needs in chronically ill children.

**Keywords** Computer-adaptive testing · Quality of life · Children · Adolescents · Reliability and validity · Self-report

## Background

Patient-reported outcomes (PRO) are increasingly used in pediatric health care and are widely administered in epidemiological as well as clinical research. Health-related quality of life (HRQoL) is an important construct for measuring the impact of a condition and its treatment on the child's and adolescent's life from their perspective. On an individual patient level, the introduction of HRQoL instruments in clinical practice facilitates communication between patient and physician and can function as screening and monitoring tools to detect and prioritize health problems and morbidities [1–3]. Thus, the patient's perspective can be incorporated in health care which is seen as an important step toward patient-centered care [4, 5]. However, the implementation of pediatric PRO measures into clinical routine has not yet advanced very far [6, 7] despite the agreement of clinicians that HRQoL measures are beneficial in the context of health care [3, 4]. Different challenges and barriers have been identified hindering implementation in clinical practice [1, 4, 8]. Besides proven validity, reliability and sensitivity to change [5, 9], aspects of suitability of use in clinical settings are demanded for HRQoL measures [8, 9]. New psychometric approaches and technological developments can contribute to address challenges and barriers.

An innovative approach for measuring HRQoL are computer-adaptive tests (CATs); CATs select and administer the most informative items from an item bank for every individual according to his or her location on the underlying latent trait [10]. The adaptive and individual selection of items provided by a CAT has some potential advantages: (1) persons with different underlying trait levels receive only those items that are especially informative and discriminating in their particular range of the trait in question, so persons will not feel overburdened compared to a static questionnaire that always contains the same items, (2) due to the characteristics of the item response theory (IRT)-based procedure, CAT scores are comparable to each other even though different respondents answered different item sets or the same person answered different item sets on different occasions, (3) in contrast to tests based on classical test theory (CTT), IRT-based measurements allow for the estimation of individual standard errors of measurement (SEs) so that one can evaluate how reliable the theta score is for a given individual, (4) item banks of a CAT may contain items with many

various types of item response formats [10] so it is possible to select the best and most informative items from different established questionnaires, (5) as other electronic tests as well, CATs provide immediate visual and numerical feedback of the results which can be handed to the clinicians or to the patients. In short, CATs are usually at least as precise as traditional paper–pencil questionnaires, but since they are tailored to each individual, fewer items need to be administered making the completion of a CAT less burdensome for respondents compared to traditional questionnaires [11].

During a prior project phase, our study group developed the Kids-CAT [12]. The development was based on the HRQoL domain structure of the well-established KIDSCREEN-27 [13]. A total of 39 instruments were carefully investigated using both qualitative (e.g., Delphi rating) and quantitative methods (e.g., confirmatory factor analyses, evaluations of item response curves, item parameter estimations). The quantitative steps include methods from both classical test theory (CTT) and item response theory (IRT). The final Kids-CAT item banks include items from 25 instruments with overall 155 items (e.g., HRQoL instruments, among others the KIDSCREEN questionnaires [14], Child Health and Illness Profile [15], Youth Quality of Life Instruments [16], and several specifically selected instruments). Further details on the development process are described elsewhere [12, 17]. The Kids-CAT covers the following five dimensions: Physical well-being (WB) (26 items), Psychological WB (46 items), Parent Relations (26 items), Social Support & Peers (26 items) and School WB (31 items). The design of the Kids-CAT items is depicted in Supplementary Fig. 1. In addition, another dimension of chronic–generic HRQoL can be linked to the Kids-CAT [18]. During the development process, the Kids-CAT dimensions were used in a simulation study [12], but the present study is the first application of the Kids-CAT in a clinical setting.

The Kids-CAT can be filled out through all devices with Internet access. The graphic design of the tool is very suitable for children/adolescents since their opinions were considered during the development of the layout. Recent work focused on the feasibility and the acceptance of the Kids-CAT from the perspective of patients and pediatricians. It was shown that the Kids-CAT is easy to complete and well accepted by patients as well as clinicians [19]. The time needed for completing the Kids-CAT was on average 7:46 min<sup>1</sup> for all participants, with adolescents needing significantly less time than younger children [19].

The aims of this paper are (1) to present the design and methods of the longitudinal Kids-CAT study and (2) to evaluate its reliability and validity in a clinical setting.

<sup>1</sup> The mean completion time refers to the time needed for the five generic core dimensions of the Kids-CAT and the additional chronic–generic dimension.

## Methods

### Study design

The Kids-CAT study is a prospective longitudinal observational study. The Kids-CAT was applied at eight measurement points over the course of one year in a cohort of children/adolescents with chronic conditions. We conducted monthly assessments in the first half year of the study. Three of these were major measurement points and were completed at the clinic (baseline, after 3, and 6 months) including the Kids-CAT and several instruments for validation purposes. As depicted in Fig. 1, four short assessments were conducted at the children's or adolescents' homes (1, 2, 4, and 5 months after baseline). After 12 months, the last major measurement point was conducted at the clinic. For all assessments, a combination of online surveys and paper–pencil questionnaires was applied. Data were collected from June 2013 to October 2014.

At the clinic, data were gathered from children/adolescents (self-report) during their scheduled clinical visits. If they had difficulties with the online survey, a study nurse provided technical support. A secure Internet data connection (https) was used. All data were collected pseudonymized with a unique identification code for each child consisting of letters and numbers. Further, their parents and the attending pediatrician filled out paper–pencil questionnaires (proxy reports). For the four short home assessments, children/adolescents received an e-mail containing a link to the Kids-CAT; parents were asked to fill out a short paper–pencil questionnaire that they had received via mail. More details on the data collection are reported elsewhere [19].

The Kids-CAT study was approved by the universities' ethics committees (Kiel and Lübeck), the Chamber of Psychotherapists in Hamburg, Germany, and the data protection office.

### Participants

Participants were consecutively recruited at two specialist outpatient departments at the University Medical Center Schleswig-Holstein in Kiel and Lübeck, located in northern Germany. Each clinic had a team with a study nurse and a group of pediatricians specialized in pulmonology, diabetology or rheumatology. We included children (7–11 years) and adolescents (12–17 years of age) with a diagnosis of bronchial asthma (J45), diabetes mellitus (E10–11) or juvenile arthritis (M08) based on the International Classification of Diseases-10 (ICD-10). Both parents and child/adolescent had to enroll in the study together and

had to be able to speak and read German. Children/adolescents and parents had to give written informed consent.

## Variables and instruments

### Kids-CAT

The latent trait estimate of each HRQoL dimension ( $\theta$ ) was calculated with the CAT algorithm using the expected a posteriori (EAP) method [20]. The CAT stops the administration of each dimension after a predefined measurement precision of a standard error ( $SE$ ) of  $\leq .32$  (which corresponds to a reliability of  $r = .90$ ) or after a maximum of seven items was administered. The minimum number of items was set to three. The maximum of seven items was based on the results of a simulation study which demonstrated that the Kids-CAT achieved a measurement precision between  $SE = .25$  and  $.50$  with on average seven items [12]. The minimum of three items per dimension was preset to avoid confusion in children in case when only one or two items had to be administered to achieve the predefined measurement precision. Items included in the Kids-CAT item banks were taken from well-established German versions of standardized questionnaires. Original phrasings, recall periods and response options were kept; only slight modifications were needed for 16 out of overall 155 items to avoid confusions in respondents answering the Kids-CAT (e.g., replacing “children” with “children and adolescents”). After calibration, item banks of the Kids-CAT were normed based on data of the German KIDSCREEN sample ( $n = 1678$ ; age:  $M = 12.65$ ;  $SD = 2.738$ ; 51.3% female) using a stratification by age (7–11 vs. 12–17 years old) and gender. More details on the development of the Kids-CAT are already published [12].

### Validation instruments

An overview of the self- and proxy instruments used in the Kids-CAT study is provided in Supplementary Tables 1 and 2. The original recall periods of the respective instruments were used, if not declared otherwise. We utilized already available and well-established German versions of the following standardized instruments to assess HRQoL and related constructs for the purpose of validating the Kids-CAT (for Cronbach's  $\alpha$  as a measure of internal consistency of the validation instruments see Supplementary Table 3):

#### *Pediatric quality of life inventory (PedsQL)*

The PedsQL 4.0 self-reported core scales contain 23 items assessing four dimensions of generic HRQoL (Physical,

**Fig. 1** Design of the Kids-CAT study. Study design with the four clinical assessments T1, T4, T7 and T8 and the four home assessments. ● = assessment at clinic; ■ = assessment at home

Month of study participation		0	1	2	3	4	5	6	7	8	9	10	11
		T1	T2	T3	T4	T5	T6	T7					T8
Child/adolescent	CAT and health items	●	■	■	●	■	■	●					●
	Validation instruments	●			●			●					●
Parent (proxy versions)	Validation instruments	●			●			●					●
	Health items			■	■			■	■				
Pediatrician	Clinical data	●			●			●					●

Emotional, Social, and School Functioning) in children/adolescents [21]. Additionally, three global sum scores can be calculated (total score, physical and psychological health summary scores). Individuals respond on a five-point rating scale from 0 = “never” to 4 = “almost always”. Reverse scoring of the items results in higher scores indicating higher HRQoL. In contrast to the original questionnaire, we used a recall period of one week. Since we used the PedsQL to validate the Kids-CAT, comparable recall periods of both measurement instruments were desirable. The German version of the PedsQL has a good internal consistency and is able to discriminate between patients with different conditions [22].

#### KIDSCREEN-27

The KIDSCREEN-27 assesses generic HRQoL and consists of 27 items measuring five subscales (Physical WB, Psychological WB, Autonomy & Parent Relation, Social Support & Peers, and School Environment). Items are answered on five-point rating scales (response options ranging from 0 = “excellent” to 4 = “poor,” 0 = “never” to 4 = “always” or 0 = “not at all” to 4 = “extremely”) with higher scores indicating higher HRQoL. All KIDSCREEN-27 subscales have robust psychometric properties including satisfactory internal consistency and test–retest reliability [13].

#### EQ-5D-Y

The EQ-5D-Y is a generic HRQoL measure with five items and a visual analog scale (VAS) [23]. We exclusively used the VAS as a measure of global health asking children/adolescents to rate their own health between 0 (“worst health state”) and 100 (“best health state you can imagine”). Recent results on convergent validity suggest that the VAS incorporates aspects of physical as well as psychological health [24].

#### General Health Item

To assess the general health status of the patients, we used the first item of the global health 10-item self-reported

PROMIS instrument [25]. The item “In general, would you say your health is:” is rated on a five-point response scale ranging from 1 = “excellent” to 5 = “poor”.

#### Global health change score

This item stems from the Short Form Health Survey (SF-36) and assesses the current perceived health status compared to four weeks ago [26]. The recall period of this item was changed from one year to four weeks to meet the needs of our study design. The item “Compared to four weeks ago, how would you rate your health in general now?” is answered on a five-point rating scale ranging from 1 = “much better now” to 5 = “much worse”.

#### The General Self-Efficacy (GSE) Scale

The ten-item GSE scale is a self-report instrument assessing a general sense of perceived self-efficacy on a four-point rating scale ranging from 1 = “not at all true” to 4 = “very true” [27, 28]. A higher sum score indicates higher self-efficacy. Good factorial and criterion validity was found [28] as well as a good internal consistency [29].

#### Students’ Life Satisfaction Scale (SLSS)

The SLSS is a seven-item self-report instrument to assess global life satisfaction in children/adolescents [30] based on the Satisfaction With Life Scale (SWLS) [30, 31]. Items of the SLSS are rated on a six-point rating scale ranging from 1 = “strongly disagree” to 6 = “strongly agree”. Higher sum scores reflect a higher satisfaction with life. Good reliability and validity have been demonstrated [30, 32].

#### Family Assessment Device (FAD)

The FAD assesses family functioning and includes seven subscales [33]. In our study only, the subscale overall Family Functioning was assessed with 12 items rated on a four-point rating scale (1 = “strongly agree” to 4 = “strongly disagree”). Lower scores indicate better family functioning.

Internal consistency of this subscale has been shown to be very good [34] to excellent [33]. Test–retest reliability and concurrent validity were demonstrated in a study with a relatively small sample [35].

#### *Social Support Scale (SSS)*

The assessment of social support was based on the German translation of the Social Support Survey [36]. The wording of items was slightly modified to be suitable for children/adolescents self-report; items not applicable to children/adolescents were deleted. The resulting eight-item shortened version assessed how often particular types of support were available to the respondent and are answered on a five-point rating scale (1 = “none of the time” to 5 = “all of the time”). A higher sum score represents more social support.

#### **Socio-demographic variables**

##### *Socioeconomic status (SES)*

The SES was operationalized by a Germany-wide well-established SES index [37, 38]. This seven-item index which has been developed on the theoretical basis of the Winkler index [39] includes the three dimensions: *education* based on the international CASMIN classification [40], *occupation* based on the International Socio-Economic Index of Occupational Status (ISEI-index) [41] and *income* operationalized in accordance with the German and European social reporting using the equivalent income. For each dimension, a score was calculated ranging from 1 to 7 points, i.e., the metric total SES score was formed by summing the three dimensions equally weighted. It ranges from 3 to 21 points. Higher scores indicate a higher SES. Additionally, a categorization by quintiles into three groups with low (first quintile), medium (second to fourth quintiles) and high (fifth quintile) SES was used. We used the cutoff points of a reference sample from the German Health Interview and Examination Survey for Children and Adolescents (KiGGS study) [38].

##### *Migration background*

Following Schenk et al. [42], migration background was defined as present if at least one of the following criteria was met: both parents were born in a country other than Germany, child and at least one parent were born in a country other than Germany or native language of the child/adolescent was not German.

#### **Statistical analysis**

Analyses reported here are based on the patient self-reported data gathered at the clinic at baseline measurement point T1. For the investigation of test–retest reliability, measurement point T2 was used additionally (self-reported data gathered at patients’ home).

#### **Reliability**

Each EAP estimate from the different CAT dimensions comes with a standard error of measurement (SE), reflecting *measurement precision*. This differs across the latent trait with extreme values being less precise [10]. We report the mean, standard deviation, and range of the SE for each of the five dimensions at measurement point T1. The mean SEs were translated into reliability coefficients using the following formula:  $r = 1 - SE^2$  to allow comparisons to measures constructed under CTT.

*Test–retest reliability* was assessed in a subsample of 97 children/adolescents at two measurement points that were approximately four weeks apart (T1 and T2). To evaluate test–retest reliability, we selected those patients with no change in their health status between both measurement points according to the self- as well as parent-reported global health change score. Intraclass correlation coefficients ( $ICC(3,1)$ ) and 95% confidence intervals were calculated between the five Kids-CAT dimensions at both measurement points [43]. A coefficient of .7 or higher was considered as adequate test–retest reliability for group comparisons [44, 45]. For evaluating individuals, a coefficient of at least .9 should be reached [46].

#### **Validity**

We calculated *intercorrelations* among Kids-CAT dimensions to evaluate *internal validity*. Concerning *convergent validity*, we expected moderate to strong Pearson correlations between the Kids-CAT dimensions and corresponding subscales of the well-established HRQoL measures PedsQL, KIDSCREEN-27 and questionnaires assessing related constructs (general health, general self-efficacy, life satisfaction, family functioning and social support). Regarding *discriminant validity*, we expected significantly lower Pearson correlations of Kids-CAT dimensions with non-corresponding subscales of the PedsQL, KIDSCREEN-27 and with non-related constructs compared to related constructs. To account for the partial overlap of items from the KIDSCREEN-27 used also in the Kids-CAT, we additionally analyzed the corrected correlations between the Kids-CAT dimensions and the KIDSCREEN-27 dimensions for items used in the Kids-CAT.

**Table 1** Sample characteristics of participants ( $N = 270$ )

	Total ( $N = 270$ )	Children; 7–11 years ( $n = 99$ )	Adolescents; 12–17 years ( $n = 171$ )
Age $M$ (SD)	12.62 (2.75)	9.53 (1.16)	14.42 (1.55)
Sex $n$ (%)			
Female	129 (47.8)	45 (45.5)	84 (49.1)
Male	141 (52.2)	54 (54.5)	87 (50.9)
Condition $n$ (%)			
Asthma	52 (19.3)	23 (23.2)	29 (17.0)
Diabetes	182 (67.4)	66 (66.7)	116 (67.8)
Juvenile arthritis	36 (13.3)	10 (10.1)	26 (15.2)
Clinic $n$ (%)			
Lübeck	108 (40.0)	39 (39.4)	69 (40.4)
Kiel	162 (60.0)	60 (60.6)	102 (59.6)
Socioeconomic status			
$M$ (SD, $n$ )	13.52 (2.97, 212)	13.49 (2.96, 82)	13.54 (2.99, 130)
Low $n$ (%)	7 (3.3)	2 (2.4)	5 (3.8)
Medium $n$ (%)	148 (69.8)	56 (68.3)	92 (70.8)
High $n$ (%)	57 (26.9)	24 (29.3)	33 (25.4)
Migration background $n$ (%)			
No	242 (94.9)	87 (94.6)	155 (95.1)
Yes	13 (5.1)	5 (5.4)	8 (4.9)

$M$  mean,  $SD$  standard deviation

Correlations of  $.5 > r < .7$  were interpreted as moderate, whereas correlations of  $.7 > r < .9$  were interpreted as high [47].

## Results

### Sample description

A total of 397 patients were asked for participation and 312 (78.6%) enrolled in the study. The computerized version of the Kids-CAT had not been available at the first weeks of the study; thus, CAT analyses presented herein are based on a sample of  $N = 270$ . For the comparison of participants and non-participants, we used the complete sample of  $N = 312$  patients. Comparing participants and non-participants we found no significant differences regarding age, sex, disease and clinic.

There were several reasons for declining participation. Concerning patients, a total of 30 (35.3%) did not wish to participate. Further reasons for non-participation were lack of time ( $n = 8$ ; 9.4%) and health reasons ( $n = 6$ ; 7.1%). Lack of sufficient reading abilities were reported by four patients (4.7%). One child (1.2%) reported a lack of sufficient abilities to deal with the Internet, and one patient (1.2%) had no Internet access at home. Parents refused participation due to lack of time ( $n = 10$ ; 11.8%),

having no interest in the study ( $n = 8$ ; 9.4%), that their child did not wish to participate ( $n = 5$ ; 5.9%), and a lack of reading abilities of their child ( $n = 4$ ; 4.7%). Concerns regarding data security ( $n = 1$ ; 1.2%) and the feeling that participation would be too burdensome ( $n = 1$ ; 1.2%) were infrequent reasons for non-participation.

Socio-demographic data of the sample used for the validation of the Kids-CAT are shown in Table 1. Patients' mean age was 12.6 years ( $SD = 2.75$ ) and 129 participants (47.8%) were female.

Figure 2 displays the distributions of HRQoL scores of the participants assessed with the Kids-CAT. By means of the Kids-CAT, HRQoL was measured in theta scores with lower thetas representing lower HRQoL. Theta of 0 represents the mean value, and theta of 1 represents the standard deviation of the reference sample used in the development phase of the Kids-CAT [12]. The median of our sample was around the norm value of the reference sample for the dimensions Physical WB and Psychological WB and slightly above the norm value for the remaining three dimensions.

### Measurement precision

The mean  $SEs$  of the Kids-CAT dimensions ranged between .38 and .49, which translates into a reliability of .86 and .76, respectively. Figure 3 shows the measurement

precision over the continuum of the latent trait for the measurement point T1. The average reliability was  $r = .80$  across the whole latent trait. The Kids-CAT measured most reliably in the lower HRQoL range of the latent trait (theta  $-2$  to  $0$ ), with an average reliability of  $r = .87$  (ranging from  $r = .82$  for Physical WB to  $r = .90$  for School WB). Measurement precision decreased in the higher range of HRQoL scores (theta  $0$  to  $2$ ) to  $r = .76$  (ranging from  $r = .71$  for Physical WB to  $r = .84$  for School WB).

Frequencies of achieved reliabilities are displayed in Table 2 including all Kids-CAT assessments at T1. The *predetermined* level of measurement precision of  $SE \leq .32$  (corresponding to a reliability of  $r \geq .9$ ) was not achieved for any measurement with the dimension Physical WB, but for 27.2% of the measurements with the dimension School WB. In total, 8.8% of the measurements across all dimensions achieved a  $SE \leq .32$  and about 50% of the measurements at T1 reached a measurement precision between  $SE \leq .45$  and  $SE > .32$  (corresponding to a reliability between  $r = .8$  and  $r = .9$ ).

Not all items from the item banks were administered by the CAT algorithm; the actual presentation of available items in the item banks per dimension ranged between 39.1% (Psychological WB) and 66.7% (Physical WB). Items administered at T1 stem from 14 different questionnaires. Across the five dimensions, 47.3% of the items in the item banks were not administered at all, because they

were not selected from the CAT algorithm due to their item parameters.

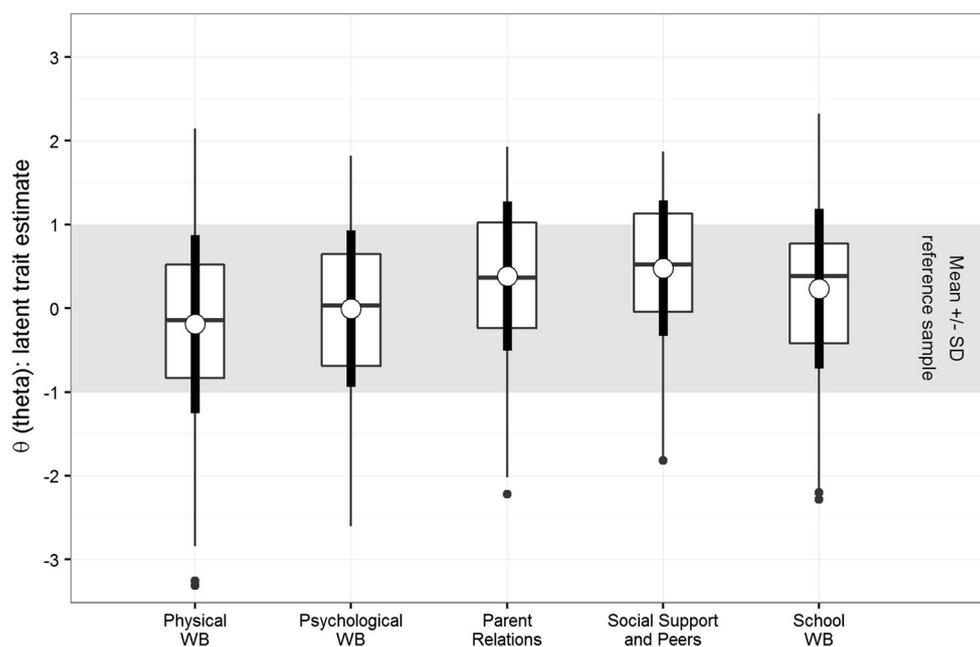
### Test–retest reliability

In the subsample of 97 patients with unchanged health status between measurement points T1 and T2, the ICCs for scale scores of both measurements ranged between  $ICC = .66$  for Physical WB and  $ICC = .77$  for School WB (Table 3).

### Validity

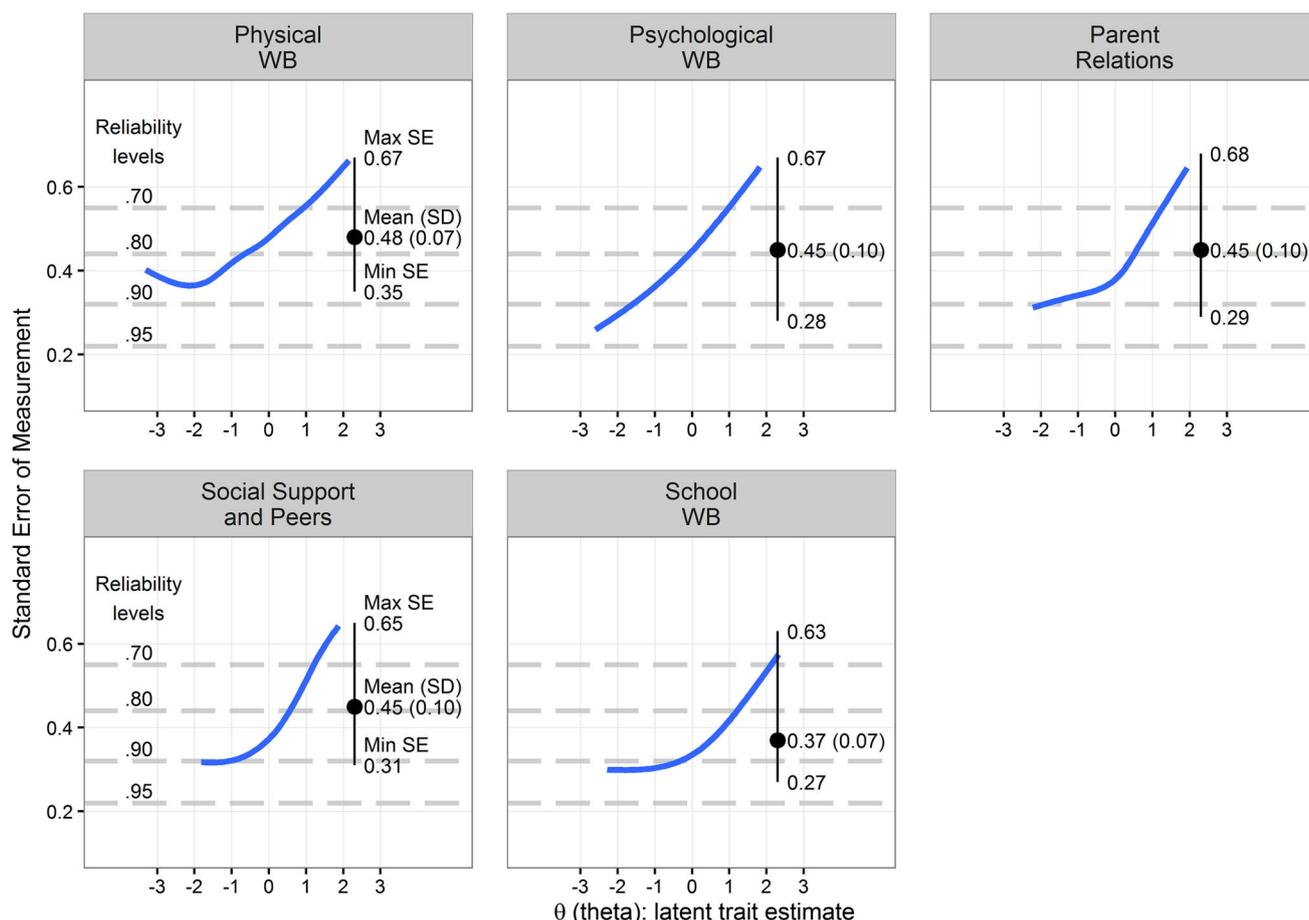
The intercorrelations as indicators of *internal validity* among the Kids-CAT dimensions ranged between  $r = .27$  for Physical WB and Social Support & Peers, and  $r = .55$  for Psychological WB and Parent Relations (see Table 4).

Results concerning convergent and discriminant validity are depicted in Table 5. The upper part of this table includes the correlations between the five Kids-CAT dimensions and the HRQoL measures. In terms of *convergent validity*, all Kids-CAT dimensions correlated moderately to highly with the corresponding dimensions of the PedsQL and the KIDSCREEN-27. However, for the Kids-CAT dimension Social Support & Peers only a low correlation was found with the PedsQL dimension Social Functioning ( $r = .33$ ).



**Fig. 2** Distribution of theta scores of the five Kids-CAT dimensions. *White circles* represent respective means, *horizontal bold lines* the medians, *vertical bold lines* the standard deviations (SD) and *hinges* the interquartile ranges (distance between the first and third quartiles). The whiskers extend from the hinges to the highest/lowest values that

are within 1.5-fold interquartile ranges. Data beyond the end of the whiskers are outliers and plotted as *small black dots*. The *underlying gray area* represents the mean plus/minus one SD of the reference sample. *WB* Well-being



**Fig. 3** Reliability graphs of the five Kids-CAT dimensions. The *x* axes show the continuum of the latent traits using theta scores, the *y* axes show the measurement precisions based on standard errors of measurement (*SE*) and reliability coefficients ( $r = 1 - SE^2$ ). The

*blue graphs* represent the particular reliability curves for each Kids-CAT dimension. The *black dots* depict the mean standard errors of measurement (standard deviations *SD* in parentheses) and the *vertical black lines* show the range of *SEs*. *WB* Well-being

**Table 2** Measurement precision and corresponding reliability levels of the five Kids-CAT dimensions (*N* = 270)

Kids-CAT dimension	<i>SE</i> > .45 <sup>a</sup> <i>n</i> (%)	.45 ≤ <i>SE</i> > .32 <sup>b</sup> <i>n</i> (%)	<i>SE</i> ≤ .32 <sup>c</sup> <i>n</i> (%)
Physical WB	162 (60.0)	108 (40.0)	0 (0)
Psychological WB	122 (45.2)	124 (45.9)	24 (8.9)
Parent Relations	118 (43.9)	138 (51.3)	13 (4.8)
Social Support & Peers	119 (44.2)	141 (52.4)	9 (3.3)
School WB	38 (14.2)	157 (58.6)	73 (27.2)

<sup>a</sup> Translates into a reliability of  $r < .80$

<sup>b</sup> Translates into a reliability of  $.80 \leq r < .90$

<sup>c</sup> Translates into a reliability of  $r \geq .90$

*Discriminant validity* was demonstrated by significantly lower Pearson correlations with non-corresponding subscales of the PedsQL and the KIDSCREEN-27 compared to corresponding subscales (bold in Table 5). For example, the correlation of the Kids-CAT dimension Physical WB with the non-corresponding dimension Social Functioning of the PedsQL ( $r = .40$ ) was significantly lower than the correlation with the corresponding PedsQL dimension

Physical Functioning ( $r = .72, z = -5.52, p < .001$ ). The only exception was the correlation between the Kids-CAT dimension Social Support & Peers with the corresponding dimension of the PedsQL, Social Functioning. Although the differences between the corresponding and non-corresponding correlations were statistically significant ( $z = -2.18, p < .05$ ), the correlation of  $r = .33$  was not very high.

**Table 3** Test–retest reliability

Kids-CAT dimension	ICC two way mixed (95% CI) $n = 97$
Physical WB	.66** (.53 to .76)
Psychological WB	.67** (.53 to .77)
Parent Relations	.73** (.61 to .81)
Social Support & Peers	.67** (.53 to .77)
School WB	.77** (.77 to .84)

WB Well-being

\*\*  $p < 0.001$ 

In additional analyses (see Supplementary Table 4), we corrected the correlations between the Kids-CAT dimensions and the KIDSCREEN-27 dimensions to account for the overlap of some KIDSCREEN-27 items used in the Kids-CAT item banks. Correlations did not change substantially.

The lower part of Table 5 depicts convergent and discriminant validity of the five Kids-CAT dimensions and related constructs. As expected, the General Health Item (reversely scored:  $r = -.63$ ) and the VAS of the EQ-5D-Y ( $r = -.60$ ) correlated higher with the dimension Physical WB than with other Kids-CAT dimensions. The SLSS had its highest correlation with the Kids-CAT dimension Psychological WB as expected ( $r = .65$ ), but correlated also moderately with Parent Relations ( $r = .54$ ). As expected, the GSE correlated moderately with the dimension Psychological WB ( $r = .51$ ), but also with School WB ( $r = .52$ ). The FAD correlated moderately (reversely scored:  $r = -.66$ ) with Parent Relations; all other correlations were substantially lower. Finally, the SSS correlated highest with Social Support & Peers ( $r = .69$ ) but not with other Kids-CAT dimensions.

## Discussion

This study aimed at presenting the design of the longitudinal Kids-CAT validation study and its reliability and validity in a sample of chronically ill children and

adolescents under real life settings. The Kids-CAT project is the first longitudinal CAT project for measuring HRQoL of children and adolescents via CAT technology in Europe.

This study showed that the *measurement precision* of the Kids-CAT was high in the area of lower HRQoL. This finding makes the tool a promising instrument for screening purposes to detect children and adolescents with impaired HRQoL. A reliable measurement of low HRQoL scores is a prerequisite for identifying children and adolescents in need for further assessment and/or treatment. Measurement precision curves found in this study were very similar to those found in a recent simulation study on the Kids-CAT [12]. The main difference between both studies was that the simulees in the simulations had lower HRQoL scores compared to our sample. For higher HRQoL scores, we found lower measurement precision of the Kids-CAT compared to lower HRQoL scores which might be due to the items of the 25 original questionnaires that constitute the item banks [12] as these were mostly problem oriented and mainly had a clinical focus. The differentiation in very high ranges of the respective HRQoL dimension is difficult and might uncover an underlying conceptual question: How can very high levels of HRQoL be differentiated at all without changing the construct from HRQoL to happiness for example?

To improve the precision in the area of higher HRQoL, there are several options. For example, the Kids-CATs stopping rule could be changed by increasing the predefined maximum number of items to be administered since seven items may not be enough to gain a satisfactory level of measurement precision across the whole continuum of the latent trait. The downside of this may be a higher respondent burden. Another option is to extend the item banks with items differentiating better in higher HRQoL ranges. However, if the primary use of the Kids-CAT is screening for children with low HRQoL who may be in need of further support or treatment, the most important feature of the Kids-CAT is to measure precisely in lower areas of the latent trait. However, the bottom line is that this study showed that the Kids-CAT strength lies exactly

**Table 4** Intercorrelations between Kids-CAT dimensions

	Physical WB $r$ ( $n$ )	Psychological WB $r$ ( $n$ )	Parent Relations $r$ ( $n$ )	Social Support & Peers $r$ ( $n$ )	School WB $r$ ( $n$ )
Physical WB	1	.53** (270)	.34** (269)	.27** (269)	.37** (268)
Psychological WB		1	.55** (269)	.29** (269)	.49** (268)
Parent Relations			1	.28** (269)	.47** (268)
Social Support & Peers				1	.32** (268)
School WB					1

WB Well-being

\*\*  $p < 0.001$

**Table 5** Convergent and discriminant validity of the five Kids-CAT dimensions with subscales of the PedsQL, KIDSCREEN-27 and other related constructs

Kids-CAT dimension	Physical WB	Psychological WB	Parent Relations	Social Support & Peers	School WB
HRQoL					
PedsQL					
Physical Functioning	<b>.72** (264)</b>	.46** (264)	.23** (264)	.16* (264)	.29** (263)
Emotional Functioning	.48** (264)	<b>.65** (264)</b>	.41** (264)	.17** (264)	.38** (263)
Social Functioning	.40** (262)	.34** (262)	.27** (262)	<b>.33** (262)</b>	.30** (261)
School Functioning	.46** (262)	.41** (262)	.29** (262)	.15* (262)	<b>.52** (261)</b>
KIDSCREEN-27					
Physical WB	<b>.77** (257)</b>	.48** (257)	.24** (256)	.21** (256)	.29** (256)
Psychological WB	.58** (259)	<b>.69** (259)</b>	.43** (258)	.27** (258)	.46** (258)
Autonomy & Parent Relations	.42** (252)	.51** (252)	<b>.57** (251)</b>	.23** (251)	.41** (251)
Social Support & Peers	.32** (260)	.28** (260)	.21** (259)	<b>.61** (259)</b>	.22** (259)
School Environment	.40** (255)	.50** (255)	.39** (254)	.22** (254)	<b>.70** (254)</b>
Related constructs					
EQ-5D-Y VAS	<b>.60** (254)</b>	.43** (254)	.27** (253)	.15* (253)	.35** (252)
General Health Item	−.63** (265)	−.51** (265)	−.33** (264)	−.15* (264)	−.33** (264)
General Self-Efficacy (GSE)	.42** (169)	<b>.51** (169)</b>	.42** (169)	.44** (169)	.52** (169)
Students' Life Satisfaction Scale (SLSS)	.45** (169)	<b>.65** (169)</b>	.54** (169)	.40** (169)	.47** (169)
Family Assessment Device (FAD)	−.20* (161)	−.41** (161)	−.66** (161)	−.22** (161)	−.29 (161)
Social Support Scale (SSS)	.23** (169)	.22** (169)	.27** (169)	<b>.69** (169)</b>	.20** (169)

Convergent validity in bold

WB Well-being, VAS visual analog scale

\*  $p < 0.01$ ; \*\*  $p < 0.001$

where its development was targeted at: to make screening of children with low HRQoL precise, short and efficient for implementation in busy clinical practice.

In the present sample of chronically ill children, approximately only half of the items in the five item banks were administered. This is due to the item selection algorithm that consistently selects the item with the highest information at the expected theta score; items never administered did not fulfill this criterion in the current sample of chronically ill children and adolescents. In future studies with children and adolescents with other conditions and without a chronic condition, we will analyze whether the same items are not administered as in the current study with patients with asthma, diabetes or juvenile arthritis.

Overall, *test–retest reliability* of the Kids-CAT was not fully satisfactory. The coefficients were above the criterion of .7 for two of the five Kids-CAT dimensions, while the remaining three dimensions were close to .7. Our results are promising concerning the use of the Kids-CAT in group comparisons, but question its use for individual comparisons according to insufficient test–retest reliability for this purpose. However, test–retest reliability was investigated over a rather long period of time in the present study (four weeks instead of two weeks as usual) and the small sample

size of 97 patient with unchanged health status could also account for these results. Future studies should thus re-investigate the test–retest reliability of the Kids-CAT in a greater sample and with a shorter time period between both measurement points. Moreover, we used different contexts at baseline (at the clinics with technical support from study nurses) versus at the short assessment one month later (respondents answering on their own at home). Although the literature found no significant differences between clinic and home assessment [48], it may be interesting to investigate potential context effects on self-reported HRQoL assessed by means of the Kids-CAT more detailed in future studies.

Intercorrelations of the Kids-CAT dimensions showed evidence for the *internal validity of the tool* since the dimensions were not too closely related to each other. *Convergent validity* was demonstrated by moderate to high correlations between the Kids-CAT dimensions and corresponding subscales of other well-established HRQoL instruments (KIDSCREEN-27 and PedsQL) and other related constructs. However, the Kids-CAT dimension Social Support & Peers correlated unexpectedly low with the PedsQL Social Functioning scale ( $r = .33$ ). This may be due to different conceptualizations of the two

dimensions. The Kids-CAT dimension focuses on feelings and perceptions concerning a broader concept of social support and the relationship to peers, whereas the PedsQL dimension focuses mainly on two narrow aspects, namely social problems with others (e.g., feeling bullied) and the comparison of social competence of the child with other children. In contrast, the Kids-CAT dimension Social Support & Peers intentionally excluded assessing “specific negative interactions like bullying tactics” [12]. Further, we found a relatively low correlation between the Kids-CAT dimension Parent Relation and the KIDSCREEN-27 dimension Autonomy & Parent Relation ( $r = .57$ ), which is most likely due to the fact that the Kids-CAT does not include the autonomy aspect of the KIDSCREEN-27 dimension. During the Kids-CAT development process we excluded items referring to the autonomy of children and adolescents due to the strict unidimensionality assumption, required in the IRT-based item bank development [12]. As expected, *discriminant validity* of the Kids-CAT was shown by lower correlations with non-corresponding dimensions of HRQoL instruments compared to corresponding dimensions.

### Limitation

Our study has some limitations. (1) The duration of data assessment at the major measurement points (T1, T4, T7 and T8) was quite long; this may concern especially young children with reading skills still developing. Since the Kids-CAT was administered at the beginning of each session, fatigue effects are not expected for Kids-CAT scores, but they cannot be excluded concerning the data assessment for some validation instruments that were assessed after the Kids-CAT. (2) Particularly the youngest children may have had difficulties reading and understanding all of the items, especially at the first assessment, as some of them were only seven years of age. (3) Since the KIDSCREEN-27 domain structure was used to build this new CAT and also some KIDSCREEN items were included in the final item banks, the validation of the Kids-CAT with the KIDSCREEN-27 is somewhat circular. To account for this limitation, we calculated also the convergent and discriminant validity between the Kids-CAT dimensions and the KIDSCREEN-27 dimensions corrected for items used in the Kids-CAT. With these analyses, we showed that the correlations did not change substantially. (4) The rather long duration of four weeks between the measurement points T1 and T2 for the investigation of the test–retest reliability and the small sample size of 97 children/adolescents with an unchanged health status limit the explanatory power of this analysis. (5) To harmonize the recall periods of measures used in the present study, we changed the recall period of the standardized PedsQL

questionnaire which may have affected its psychometric properties to some extent. (6) Since we had a convenience sample, families with a low SES and patients with migration background were underrepresented in our study. This restricts the generalizability of our results. Currently, the Kids-CAT has been administered to a German nationwide representative sample and the question whether the Kids-CAT provides reliable and valid results also in children with a low SES and/or migration background can be investigated soon.

### Implications

This study is part of the large longitudinal Kids-CAT study including chronically ill children and adolescents. Future publications will investigate the longitudinal performance of the Kids-CAT and compare its performance between different disease groups. Furthermore, we will report data of large healthy norm samples of the general population. Results of our feasibility study [19] in combination with the present findings regarding reliability and validity of the Kids-CAT lead to the conclusion that the Kids-CAT is a short, feasible, precise, reliable and valid instrument for children and adolescents with chronic illnesses (such as asthma, diabetes and juvenile arthritis). It is a computerized version with a child-friendly graphic design, and user-friendly reports and can be used by clinicians and researchers for screening of children with impaired HRQoL and monitoring HRQoL in children during routine treatment. Although in the present study only two children out of 397 declined participation due to lack of Internet skills or access, the fact that the Kids-CAT relies on the Internet has to be considered in terms of potential disadvantage of children from more deprived socioeconomic regions or countries. Future research will need to explore whether the use of the Kids-CAT facilitates the communication about psychosocial problems between patients and clinicians and whether it is useful in detecting and prioritizing health problems. The Kids-CAT has the potential to become an important tool to strengthen patient-centeredness in pediatric health care.

**Acknowledgements** We thank all children, adolescents and their parents, who participated in the Kids-CAT study. We are grateful to the highly motivated study nurses Anja Bünte and Andrea Knaak and all pediatricians who contributed to the Kids-CAT project. We thank all student assistants and interns for their contributions to data entry and management. We thank the Federal Ministry of Education and Research for funding this research project. The Kids-CAT Study Group comprises: A. Bünte (Department of Pediatrics, University Medical Center Schleswig-Holstein, Campus Kiel, Arnold-Heller-Straße 3, House 9, 24105 Kiel, Germany), K. Gulau (Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics, Research Unit Child Public Health, Center for Psychosocial Medicine, University Medical Center Hamburg-Eppendorf,

Martinstraße 52, 20246 Hamburg, Germany), A. Knaak (Hospital for Pediatrics and Adolescent Medicine, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany), H. Muehlan (Department Health & Prevention, Ernst-Moritz-Arndt University; Robert-Blum-Straße 13, 17487 Greifswald, Germany), S. Schmidt (Department Health & Prevention, Ernst-Moritz-Arndt University; Robert-Blum-Straße 13, 17487 Greifswald, Germany) and S. v. Sengbusch (Hospital for Pediatrics and Adolescent Medicine, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany).

**Funding** This study was funded by the German Federal Ministry of Education and Research (grant number 0010-01GY1111).

#### Compliance with ethical standards

**Conflict of interest** D. Barthel, C. Otto, S. Nolte, A.-K. Meyrose, F. Fischer, J. Devine, O. Walter, A. Mierke, K. I. Fischer, U. Thyen, M. Klein, T. Ankermann, M. Rose, U. Ravens-Sieberer, A. Bünte, K. Gulau, A. Knaak, H. Muehlan, S. Schmidt and S. v. Sengbusch declare that they have no conflict of interest.

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

**Informed consent** Informed consent was obtained from all individual participants included in the study.

## References

- Greenhalgh, J. (2009). The applications of PROs in clinical practice: What are they, do they work, and why? *Quality of Life Research*, 18(1), 115–123.
- Valderas, J., Kotzeva, A., Espallargues, M., Guyatt, G., Ferrans, C., Halyard, M. Y., et al. (2008). The impact of measuring patient-reported outcomes in clinical practice: A systematic review of the literature. *Quality of Life Research*, 17(2), 179–193.
- Varni, J. W., Burwinkle, T. M., & Lane, M. M. (2005). Health-related quality of life measurement in pediatric clinical practice: An appraisal and precept for future research and application. *Health and quality of life outcomes*, 3(1), 1.
- Nelson, E. C., Eftimovska, E., Lind, C., Hager, A., Wasson, J. H., & Lindblad, S. (2015). Patient reported outcome measures in practice. *BMJ*, 350, g7818.
- Dawson, J., Doll, H., Fitzpatrick, R., Jenkinson, C., & Carr, A. J. (2010). The routine use of patient reported outcome measures in healthcare settings. *BMJ*, 340, c186.
- 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(1), 66.
- 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(4), 742–764.
- Kroenke, K., Monahan, P. O., & Kean, J. (2015). Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. *Journal of Clinical Epidemiology*, 68(9), 1085–1092.
- Higginson, I. J., & Carr, A. J. (2001). Using quality of life measures in the clinical setting. *BMJ*, 322(7297), 1297–1300.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum Associates.
- 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.
- Devine, J., Otto, C., Rose, M., Barthel, D., Fischer, F., Mühlhan, H., et al. (2015). A new computerized adaptive test advancing the measurement of health-related quality of life (HRQoL) in children: the Kids-CAT. *Quality of Life Research*, 24(4), 871–884.
- Ravens-Sieberer, U., Herdman, M., Devine, J., Otto, C., Bullinger, M., Rose, M., et al. (2014). 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(3), 791–803.
- Ravens-Sieberer, U., & The European KIDSCREEN Group. (2006). *The KIDSCREEN questionnaires—quality of life questionnaires for children and adolescents—handbook*. Lengerich: Pabst Science Publ.
- Starfield, B., Riley, A., Forrest, C., Green, B., Robertson, J., & Rajmil, L. (2007). *Child health and illness profile (CHIP). A comprehensive assessment of health and functioning of children and adolescents*. Baltimore: Johns Hopkins Bloomberg School of Public Health.
- Topolski, T., Edwards, T., & Patrick, D. (2002). *User's manual and interpretation guide for the youth quality of life (YQOL) instruments*. Seattle: University of Washington.
- Devine, J., Otto, C., Rose, M., Barthel, D., Fischer, F., Mühlhan, H., et al. (2015). Erratum to: A new computerized adaptive test advancing the measurement of health-related quality of life (HRQoL) in children: the Kids-CAT. *Quality of Life Research*, 24(9), 2303.
- Muehlan, H., Schmidt, S., Devine, J., Walter, O., Fischer, F., Nolte, S., et al. (2014). Assessing disease-related quality of life in children and adolescents with chronic conditions: development and application of a chronic-generic CAT derived from the DISABKIDS framework. *Quality of Life Research*, 23, 4.
- Barthel, D., Fischer, K., Nolte, S., Otto, C., Meyrose, A.-K., Reisinger, S., et al. (2016). Implementation of the Kids-CAT in clinical settings: a newly developed computer-adaptive test to facilitate the assessment of patient-reported outcomes of children and adolescents in clinical practice in Germany. *Quality of Life Research*, 25, 585–594.
- Bock, R. D., & Mislavy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444.
- Varni, J. W., Seid, M., & Kurtin, P. S. (2001). PedsQL™ 4.0: Reliability and validity of the Pediatric Quality of Life Inventory™ version 4.0 generic core scales in healthy and patient populations. *Medical Care*, 39(8), 800–812.
- Felder-Puig, R., Frey, E., Proksch, K., Varni, J., Gadner, H., & Topf, R. (2004). Validation of the German version of the Pediatric Quality of Life Inventory™ (PedsQL™) in childhood cancer patients off treatment and children with epilepsy. *Quality of Life Research*, 13(1), 223–234.
- Wille, N., Badia, X., Bonsel, G., Burström, K., Cavrini, G., Devlin, N., et al. (2010). Development of the EQ-5D-Y: A child-friendly version of the EQ-5D. *Quality of Life Research*, 19(6), 875–886.
- Ravens-Sieberer, U., Wille, N., Badia, X., Bonsel, G., Burström, K., Cavrini, G., et al. (2010). Feasibility, reliability, and validity of the EQ-5D-Y: Results from a multinational study. *Quality of Life Research*, 19(6), 887–897.
- Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health

- summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research*, 18(7), 873–880.
26. Bullinger, M. (1995). German translation and psychometric testing of the SF-36 health survey: preliminary results from the IQOLA project. *Social Science and Medicine*, 41(10), 1359–1366.
  27. Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). Windsor: NFER-NELSON.
  28. Schwarzer, R., & Jerusalem, M. (1999). *Skalen zur erfassung von Lehrer-und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame schulen*. Berlin: Freie Universität Berlin.
  29. Schwarzer, R., Mueller, J., & Greenglass, E. (1999). Assessment of perceived general self-efficacy on the Internet: Data collection in cyberspace. *Anxiety Stress and Coping*, 12(2), 145–161.
  30. Huebner, E. S. (1991). Initial development of the student's life satisfaction scale. *School Psychology International*, 12(3), 231–240.
  31. Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71–75.
  32. Huebner, E. S., & Alderman, G. L. (1993). Convergent and discriminant validation of a children's life satisfaction scale: Its relationship to self-and teacher-reported psychological problems and school functioning. *Social Indicators Research*, 30(1), 71–82.
  33. Epstein, N. B., Baldwin, L. M., & Bishop, D. S. (1983). The McMaster family assessment device. *Journal of Marital and Family Therapy*, 9(2), 171–180.
  34. Byles, J., Byrne, C., Boyle, M. H., & Offord, D. R. (1988). Ontario child health study: Reliability and validity of the general functioning subscale of the McMaster Family Assessment Device. *Family Process*, 27(1), 97–104.
  35. Miller, I. W., Epstein, N. B., Bishop, D. S., & Keitner, G. I. (1985). The McMaster family assessment device: Reliability and validity. *Journal of Marital and Family Therapy*, 11(4), 345–356.
  36. Donald, C. A., & Ware, J. E. (1984). The measurement of social support. *Research in Community and Mental Health*, 4, 325–370.
  37. Lampert, T., Kroll, L. E., & Stolzenberg, H. (2013). Messung des sozioökonomischen status in der studie „Gesundheit in Deutschland aktuell“ (GEDA). *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 56(1), 131–143.
  38. Lampert, P. D. T., Mütters, S., Stolzenberg, H., Kroll, L. E., & Group, K. S. (2014). Messung des sozioökonomischen status in der KIGGS-studie. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, 57(7), 762–770.
  39. Winkler, J., & Stolzenberg, H. (1999). Der Sozialschichtindex im Bundes-Gesundheitssurvey. *Gesundheitswesen*, 61(2), S178–S183.
  40. Brauns, H., Scherer, S., & Steinmann, S. (2003). The CASMIN educational classification in international comparative research. In Hoffmeyer-Zlotnik J.H.P., & Wolf, C. (Eds.) *Advances in cross-national comparison* (pp. 221–244). New York: Springer.
  41. Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21(1), 1–56.
  42. Schenk, L., Bau, A.-M., Borde, T., Butler, J., Lampert, T., Neuhauser, H., et al. (2006). Mindestindikatorenansatz zur erfassung des migrationsstatus. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 49(9), 853–860.
  43. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420.
  44. Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284.
  45. Fayers, P., & Machin, D. (2013). *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*. Chichester: Wiley.
  46. Nunnally, J. C., & Bernstein, I. (1994). The assessment of reliability. *Psychometric Theory*, 3(1), 248–292.
  47. Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences*. Belmont: Wadsworth Publishing.
  48. Maier, A., Holm, T., Wicks, P., Steinfurth, L., Linke, P., Münch, C., et al. (2012). Online assessment of ALS functional rating scale compares well to in-clinic evaluation: A prospective trial. *Amyotrophic Lateral Sclerosis*, 13(2), 210–216.