Determining at what age children provide sound self-reports: An illustration of the validity-index approach

Conijn, J.M.; Smits, N.; Hartman, E.E.

DOI
10.1177/1073191119832655

Publication date
2020

Document Version
Final published version

Published in
Assessment

Article

Determining at What Age Children Provide Sound Self-Reports: An Illustration of the Validity-Index Approach

Assessment
2020, Vol. 27(7) 1604–1618

DOI: 10.1177/1073191119832655
journals.sagepub.com/home/asm
SAGE

Judith M. Conijn[1], Niels Smits[1], and Esther E. Hartman[2]

Abstract

In psychological assessment of children, it is pivotal to establish from what age on self-reports can complement or replace informant reports. We introduce a psychometric approach to estimate the minimum age for a child to produce self-report data that is of similar quality as informant data. The approach makes use of statistical validity indicators such as person-fit and long-string indices, and can be readily applied to data commonly collected in psychometric studies of child measures. We evaluate and illustrate the approach, using self-report and informant-report data of the PedsQL, a pediatric health-related quality of life measure, from 651 child–mother pairs. To evaluate the approach, we tested various hypotheses about the validity of the self-report data, using the G_n^p person-fit index as the validity indicator and the mother informant-data as a benchmark for validity. Results showed that G_n^p discriminated between self-reports of younger and older children, between self-reports of children that completed the PedsQL alone or with a parent, and between self-reports and informant reports. We conclude that the validity-index approach has good potential for future applications. Future research should further evaluate the approach for different types of questionnaires (e.g., personality inventories) and using different validity indices (e.g., response-bias indices).

Keywords

children, PedsQL, person-fit index, self-report, validity index

In child research, the combined use of questionnaire or interview data of multiple informants, such as parents, teachers, clinician, or child self-reports, is best practice (e.g., Dirks, De Los Reyes, Briggs-Gowan, Cella, & Wakschlag, 2012; Taber, 2010). However, for some important constructs in child research, such as perceived parenting behavior or quality of life, child reports are considered the primary source of data (e.g., Eiser & Varni, 2013; Taber, 2010; Yi-Frazier et al., 2016). In research that is primarily based on child reports, a vital question comes up: From what age on can children provide valid self-report data?

The minimum age required for producing valid self-report data depends on the specific questionnaire properties, such as the complexity of item phrasing and the level of abstractness of items (e.g., Taber, 2010). For example, overt behaviors are easier to report for children than emotions (Eddy, Khastou, Cook, & Amtmann, 2011). Various researchers have therefore argued that determining the minimum age for a given self-report measure should be a part of the construction process of a child measure (e.g., Cole et al., 2018; Eddy et al., 2011; Landgraf, van Grieken, & Raat, 2018; Solans et al., 2008; U.S. Food and Drug Administration, 2009). In this article, we propose an approach based on statistical validity indices to estimate that minimum age, using an example from health-related quality of life (HRQOL) assessment.

Common methods for evaluating the reliability and validity of children's reports are to assess Cronbach's alpha, test–retest reliability, the extent of missing item scores, correlations with parent's scores, factor structure, and known-group validity analysis (e.g., Limbers, Newman, & Varni, 2008; Tsze, von Baeyer, Bulloch, & Dayan, 2013; Zapolski & Smith, 2013). These criteria are useful sources of information, but may also provide an incomplete view on the validity of self-reports in young children. First, although Cronbach's alpha quantifies the degree that item scores are consistent with each other, many response biases that are common in children, such as extreme response style and nondifferentiation (Chambers & Johnston, 2002; Shelton, Frick, & Wootton, 1996), may increase instead of decrease alpha if all items are worded in the same direction (Peer &

[1]University of Amsterdam, Amsterdam, Netherlands
[2]Tilburg University, Tilburg, Netherlands

Corresponding Author:
Judith M. Conijn, Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands.
Email: j.m.conijn@uva.nl

Gamliel, 2011). Second, correlations with parent reports cannot be taken as suitable criteria, as no informant score can be considered an accurate gold standard to assess child experiences or disorders (De Los Reyes & Kazdin, 2005; Zapolski & Smith, 2013). Third, test–retest coefficients may underestimate reliability because true scores may change across short periods in children due to their developmental phase (Roberts & DelVecchio, 2000). Finally, if evidence is found for poor psychometric properties for a given young age group, none of these methods provide information on the types of response biases that may have occurred.

To complement these existing psychometric approaches, we propose a novel method for investigating the age at which a given self-report scale produces valid results. This "validity-index approach" uses post hoc validity indices derived from the respondent's pattern of item scores (e.g., Huang, Curran, Keeney, Poposki, & DeShon, 2012; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016). Validity indices quantify repetitive responding, inconsistent or random responding, and response biases. Based on validity-index values, response patterns can be classified as either "normal" or "suspect", the latter meaning that the resulting test score is likely to be invalid. Recent research has shown the usefulness of validity indices for detecting nonoptimal response strategies (e.g., acquiescence response bias) and inconsistent or repetitive responding in survey data (e.g., Conijn, van der Ark, & Spinhoven, 2020; Meade & Craig, 2012; Niessen et al., 2016). These indices can also be expected to detect young children's aberrant response behaviors, as research has shown that young children may produce inconsistent or repetitive response patterns due to various factors: limited cognitive ability and memory skills, limited attention span (Eddy et al., 2011; Eiser, Mohay, & Morse, 2000; Reeve et al., 2017), response biases in using rating scales such as choosing the extremes (Chambers & Johnston, 2002; Davis et al., 2007), and displaying nondifferentiation among items that assess opposite constructs (Shelton et al., 1996).

Application of the validity-index approach requires data from a self-report measure for children and a parallel informant version of this measure, and works as follows. First, data are collected from parents (or other appropriate adult informants) as well as their children from a broad age range around the excepted minimum age required for self-reporting (e.g., 4 to 10 years) on parallel self-report and informant versions of the questionnaire. Such data are commonly collected for child measures when reference values and psychometric properties are established. Second, based on the data properties, suitable validity indices are selected. Third, one or more validity indices are computed for both the self-report and the parent-report. Fourth, for separate subgroups based on the child's age, the classification rates and mean validity-index values are compared across the child and parent data. Children are expected to have poorer validity-index values compared with parents, until a certain age at which there is no substantial difference between validity-index values and classification rates across parents and children; that age is the estimated age at which the self-report is likely to produce self-report data, that is (at least), of equal psychometric quality as data produced by parents or other adult informants.

We illustrate and evaluate the validity-index approach using data of the pediatric health-related quality of life (PedsQL), a pediatric HRQOL measure. HRQOL is a subjective multidimensional construct incorporating several broad domains that can be affected by one's health status (World Health Organization, 1947), at least including perceived physical, mental, and social functioning (Aaronson et al., 1991). HRQOL should be assessed from the individual's perspective whenever possible (Eiser et al., 2000; U.S. Food and Drug Administration, 2009), following that self-reports are most appropriate in studying children's HRQOL. In the past two decades, various HRQOL self-report scales have been developed for children aged 4 to 7 years (e.g., Matza, Swensed, Flood, Secnik, & Leidy, 2004, Solans et al., 2008; Wallander & Koot, 2016). The most widely used scale is the PedsQL (Varni, Seid, & Kurtin, 2001; Varni, Seid, & Rode, 1999), with a suggested minimum age of 5 years. The ages at which children provide self-report data on HRQOL are low compared with the minimum ages used in other assessment contexts such as life satisfaction or personality. In those contexts, self-reports are usually administered from 8 years on (Burns, 2002; De Los Reyes & Kazdin, 2005; Huebner, 1991).

Research into the validity of self-report HRQOL data collected in young children is inconclusive. For example, whereas studies on the original English PedsQL univocally conclude that children aged 5 years or older can provide valid and reliable self-report data (Limbers et al., 2008; Varni, Limbers, & Burwinkle., 2007a, 2007b, 2007c), those using other language versions have found low reliability for 5- to 7-year-olds (Bastiaansen, Koot, Ferdinand, & Verhulst, 2004; Ferreira, Baltazar, Cavalheiro, Cabri, & Gonçalves, 2014; Kobayashi & Kamibeppu, 2010). Moreover, discrepancies between child and informant ratings have commonly been found larger in younger children (about 5 to 7 years) than in older children (e.g., Yi-Frazier et al., 2016). Some researchers have taken these discrepancies as an argument for administering scales to children instead of using parent informant reports (e.g., Jozefiak, 2014; Thompson et al., 2014; Yi-Frazier et al., 2016), "there is a disconnection between parents' and youths' perceptions of youth HRQOL" (Yi-Frazier et al., 2016), while others suggest that discrepancies may be an indication of children's incapability of producing reliable and valid self-report data (Woolley, Bowen, & Bowen, 2004).

Study Aim

Our aim is to evaluate and illustrate the validity-index approach by an application to PedsQL data. Specifically, we use self- and informant report data from the PedsQL version for children aged 5 to 7 years ($n = 450$ child–mother pairs) and the PedsQL version for children aged 8 to 12 years ($n = 346$ child–mother pairs). Previous studies on the psychometric properties of these PedsQL scales suggest that children aged 8 years or older can independently provide valid and reliable PedsQL self-report data (e.g., Varni et al., 2007a, 2007c). For children aged 5 to 7 years, reliability and validity of self-report PedsQL data seem to depend on the method of administration. In studies in which children were assisted by a parent to complete the PedsQL, the PedsQL showed acceptable psychometric properties (Limbers et al., 2008; Varni et al., 2007a, 2007c) but not in studies in which the children were interviewed and parents were absent (Bastiaansen et al., 2004; Ferreira et al., 2014; Kobayashi & Kamibeppu, 2010).

Based on (a) the expectation that the validity-index approach works and (b) previous research into the psychometric properties of the PedsQL and children's cognitive limitations and response biases in using rating scales, we formulated the following hypotheses:

Hypothesis 1: Validity-index values for self-reports of 5- to 7-year-olds children who completed the PedsQL separately from a parent are poorer on average than validity-index values for self-reports of 5- to 7-year-olds children who completed the PedsQL together with a parent, but these differences are smaller for older children.

Hypothesis 2: When children complete the PedsQL separately from a parent, validity-index values are poorer on average for children's self-reports than for informant-reports of their mothers, but these differences are smaller for older children.

These hypotheses test the usefulness of the validity-index approach. Confirmation of these hypotheses suggests that the validity indices can discriminate valid from non-valid response patterns. In contrast, if our data does not show the expected group-differences in validity-index values, the usefulness of the validity-index approach is questionable and should be further investigated.

In case we find support for our hypotheses, we provide an application example of the validity-index approach using the PedsQL data. Specifically, we illustrate how the approach can be used to estimate the age that a child can independently provide PedsQL self-report data with similar psychometric quality as corresponding informant data.

Method

Participants and Procedure

This study used secondary data collected from 2008 through 2011, at primary schools in the Southeast of the Netherland, by various bachelor- or master-level students supervised by the third author (E. E. H.). Only children with a basic proficiency in the Dutch language were included. Exact data on participation rates are unavailable as the number of families invited to participate was not recorded during the study by most students. Among three subsets of data with known participation rates, the rates were 0.18 ($n = 50$), 0.29 ($n = 13$), and 0.78 ($n = 123$).

The study included a young age group (4 to 8 years old; mean age = 6.1) and an older age group (7 to 13 years old; mean age = 9.7), for which the procedure was different. The young age group included 450 children of whom 209 completed the PedsQL questionnaire separately from their parents ("separate condition") and 167 completed the PedsQL questionnaire together with one of their parents ("together condition"). For the remaining 73 children, the method of administration was unknown due to coding errors. Assignment to one of the two conditions was done randomly per school class. In the separate condition, children were interviewed by a research assistant, who marked the child's answers to the matching response options of the PedsQL. Thereafter, the children were given an information letter and an informant-version of the PedsQL to take home for their parents. Parents were asked to complete the informant version within a week. Only when at least one parent returned the informant version of the questionnaire, the corresponding self-report was used for the study. The children in the together condition completed the PedsQL at home. Parents were instructed to complete the informant-version first and subsequently to assist their children to complete the self-report by reading the questions out loud and mark the answer of the child, but not to direct the child in choosing a particular answer.

The older age group included 346 children. In this subgroup, there was no "together condition" because children were assumed to read the PedsQL themselves and were instructed to complete the PedsQL separately from their parents. A package with information about the study, a set of questionnaires and a form of consent was handed out in class and taken home. In case the parents decided to participate, fathers and mothers both completed the questionnaires and signed the informed consent. When families had more than one child in the target age group, parents decided which child participated. Children and their parents were instructed to complete the questionnaire separately. In case the child did not understand an item, the parents were permitted to assist without directing the child in choosing a

particular answer. The package of questionnaires and informed consent had to be returned within 2 weeks and put in a box in the classroom, together in one envelope. For 60 children in the older age group, it was unclear whether parents had received the explicit instruction to complete the PedsQL separately from their child.

In the current study, the mother-reported data and self-report data were used, but the father reported data were excluded. This approach is representative to common practice in child research: most often only one parent, commonly the mother, is administered an informant version of a measure (e.g., Jansen, Bodden, Muris, van Doorn, & Granic, 2017; Pereira et al., 2015).

PedsQL

The PedsQL™ 4.0 Generic Core Scales were designed to measure pediatric HRQOL in both healthy and patient populations (e.g., Varni et al., 1999). The PedsQL™ 4.0 Generic Core Scales include 23 items divided across four subscales: Physical Functioning (eight items), Emotional Functioning (five items), Social Functioning (five items), and School Functioning (five items). Higher scores indicate a poorer HRQOL. Different versions and corresponding informant versions are available for different ages. In this study, we used the PedsQL™ 4.0 Generic Core Scales for 5- to 7-year-olds and for 8- to 12-year-olds (henceforth, denoted as $PedsQL_{5\text{-}7}$ and $PedsQL_{8\text{-}12}$). For each of the two PedsQL instruments, a self-report version and an informant-version is available. In this study, the self-report $PedsQL_{5\text{-}7}$ items were rated as: 4 = never a problem, 2 = sometimes a problem, and 0 = almost always a problem, and response options are indicated with emoticons (i.e., smileys). The informant-report $PedsQL_{5\text{-}7}$ and the $PedsQL_{8\text{-}12}$ (both informant and self-report version) were rated as 4 = never a problem; 3 = almost never a problem; 2 = sometimes a problem; 1 = often a problem; 0 = almost always a problem. For computing the PedsQL total scores, missing item scores were imputed using the mean item score. Higher total scores indicate better HRQOL.

Validity Indicators

For application in the validity-index approach, we considered four validity indices quantifying different types of invalid responding: inconsistent/random responding, extreme response bias, agreement response bias, and repetitive responding. However, preliminary analyses showed that due to various data and scale properties, most validity indices were inappropriate for application to the PedsQL data. The extreme response bias index (i.e., the relative number of responses in the extreme response categories; Van Herk, Poortinga, & Verhallen, 2004) and the Long string index of repetitive responding (i.e., the maximum length of a string of identical answers; DeSimone, Harms, & DeSimone, 2015) were inappropriate due to the general population sample. The sample generally had a high HRQOL, resulting in left-skewed PedsQL item scores. When item scores are skewed, floor effects cause the Long string index to be inflated for respondents with high-trait values due to long strings of the maximum item score (Conijn, Franz, Emons, De Beurs, & Carlier, 2019). A similar problem applied to the index of extreme response bias. Floor effects cause persons with high-trait values to have many "extreme" responses. Finally, an index of agreement response bias (i.e., the tendency to agree with statements regardless of content) was inappropriate because it requires scales with both negatively and positively formulated items, or items that tap unrelated content. We therefore only used an index to quantify inconsistent/random responding. From several available validity indices designed to detect inconsistent/random responding (Meade & Craig, 2012; Niessen et al., 2016), we choose to use the G^p person-fit statistic and its normed version. Both statistics have the advantage of high power for detecting inconsistent responding compared with most other validity indices and less restrictive assumptions compared with parametric item response theory-based person-fit indices (Emons, 2008; Meijer, 1994; Niessen et al., 2016).

The G^p index is a person-fit statistic for polytomous items and counts the number of Guttman errors in a response pattern (e.g., Meijer, 1994), which can be regarded as the number of inconsistencies within a response pattern. For items with a polytomous response format, a Guttman error occurs when a respondent endorses a less popular item category (i.e., indicative of high symptom-severity) without endorsing a more popular one (i.e., indicative of low symptom-severity). Category popularity is based on the average endorsement rates across persons in the sample. The maximum value of G^p for a response pattern depends on the corresponding total score. By contrast, its normed version, G_n^p (Emons, 2008) weights the number of Guttman errors in a response pattern by its possible maximum value given the respondent's total score and ranges from 0 to 1. G^p and G_n^p both have specific advantages (Conijn et al., 2019; Emons, 2008): G^p has higher power to detect invalid response patterns and is not affected by possible bias in the respondents' total score, while G_n^p is less confounded with the total score and therefore allows for a fairer comparison of persons with a different latent trait value. Both G^p and G_n^p work best as a validity indicator when they are applied to unidimensional data. As the PedsQL has subscales measuring distinct constructs, we first computed the indices for separate subscales, and next averaged the indices into multiscale G^p and G_n^p statistics (Niessen et al., 2016). Higher values of G^p and G_n^p indicate more response inconsistencies and therefore indicate less valid response patterns.

Statistical Analyses

Appropriateness of Validity Indices. The interpretation of G^p and G_n^p as validity indicators requires the subscale data to be dominated by a single strong dimension, the subscale item scores to be locally independent given the latent dimension, and the item step response functions to be monotone increasing (Emons, 2008). For each of the PedsQL subscales, we assessed the assumption of unidimensionality by evaluating the percentage of variance in item scores explained by the first component, the ratio between the first and second eigenvalue, and using parallel analysis in the "nFactors" *R* package (Raiche, 2010). A ratio between the first and second eigenvalue of 1/3 or smaller was considered as evidence for a single strong dimension (Slocum-Gori & Zumbo, 2011) and a sufficiently strong dimension for application of a person-fit statistic (Conijn, Emons, & Sijtsma, 2014). The assumption of monotone increasing item step response functions was tested using the "mokken" *R* package (van der Ark, 2007), using $\alpha = .05$. We did not separately assess the local independence assumption because we assumed that for the short subscales used in this study, sufficient unidimensionality would also imply sufficient local independence for person-fit analysis.

Next, a choice between G^p and G_n^p should be made for use in the main analysis. The G^p index was chosen for the main analysis if there were no substantial differences in average total PedsQL scores between mothers and children and between age groups. In that case, the G_n^p index was used to replicate the analyses (i.e., the index was used in sensitivity analyses). If nontrivial differences in average total PedsQL scores were found, we used G_n^p in the main analysis, while G^p was used to replicate the analyses. Finally, we inspected the PedsQL item score distributions. If the more extreme categories were rarely endorsed, we merged different categories into a single category to obtain reliable estimates of the category endorsement rates (e.g., Niessen et al., 2016).

Computing Validity Indices and Classifications. We conducted separate analyses for the data of the $\text{PedsQL}_{5\text{-}7}$ and data of the $\text{PedsQL}_{8\text{-}12}$. For the $\text{PedsQL}_{5\text{-}7}$, the response scale of the self-report and the informant-report version was different. We therefore recoded the item scores of the informant version into the same 3-point response scale that was used in the self-report version, before computing the validity indices (score $1 \rightarrow 0$, score $3 \rightarrow 4$).

The "PerFit" *R* package (Tendeiro, Meijer, & Niessen, 2016) was used to compute the G^p and G_n^p indices, and for imputing missing item scores using a nonparametric model imputation method (Tendeiro et al., 2016). We used 20 different missing value imputations, and took the average validity-index value as the final value. Response patterns with three or more missing item scores (i.e., more than 10% missing values) were discarded from the analyses; imputing item scores for these respondents was expected to distort validity-index values too much.

Based on G^p and G_n^p, response patterns can be classified as severely inconsistent, suggesting that the resulting total score is probably invalid or "'suspect." We computed cutoff values using a bootstrap procedure based on the graded response model (i.e., an item response theory for polytomous data) with 500 replications and a one-sided α-level of .05 (e.g., Seo & Weiss, 2013; Sinharay, 2017). The latent trait values and item parameters for generating the bootstrap replications were estimated using the "mirt" R package (Chalmers, 2012), using semiparametric Davidian curves to account for a skewed distribution of the latent trait values (Woods & Lin, 2009). Respondents who had a G^p or G_n^p value larger than the cutoff value (i.e., the average cutoff value across the 500 replications) were classified as having a suspect PedsQL response pattern. The *R* code for computing bootstrap cutoff values can be obtained from the first author.

Hypotheses 1 and 2. To address Hypothesis 1, we analyzed the $\text{PedsQL}_{5\text{-}7}$ self-report data using multiple regression analyses. The selected validity index was used as the dependent variable, and the independent variables were condition (separate = 0; together = 1), age of the child (centered at the minimum observed value of age), and the interaction between age and condition. A negative main effect of condition and positive interaction effect between condition and age support our hypothesis. These effects indicate that the youngest children in the separate condition provide higher validity-index values (i.e., suggesting poorer validity) than the youngest children in the together condition, and that the negative effect of the together condition on validity-index values decreases with increasing age.

To address Hypothesis 2, we analyzed the mother-reported and self-reported data of the $\text{PedsQL}_{5\text{-}7}$ (separate condition only) and $\text{PedsQL}_{8\text{-}12}$. For each data set, we used a linear mixed effects model with the selected validity index as the dependent variable and the child–mother pairs as the grouping variable. Childs' age (centered at the observed minimum age) and informant (self-report = 0; mother report = 1) and their interaction were the independent variables. A negative main effect of informant and positive interaction effect between age and informant supports our hypothesis. These effects indicate that for the youngest children, self-reports have higher average validity-index values (and poorer validity) compared with informant reports and that the negative effect of a mother-report on validity-index values decreases with age. We used the "lme4" *R* package (Bates, Maechler, Bolker, & Walker, 2015) to estimate the linear mixed effects models.

All analyses were conducted in *R* version 3.3.5 (R Core Team, 2018). Effects were tested using $\alpha = .05$, one sided. R^2 was computed for the multiple regression model. To

Table 1. Number of Mother-Child Pairs and Average PedsQL Total Scores For Mothers and Children Grouped by Child's Age in Years.

	PedsQL$_{5\text{-}7}$ data				PedsQL$_{8\text{-}12}$ data						
	5 years	6 years	7 years	8 years	7 years	8 years	9 years	10 years	11 years	12 years	13 years
	Number of mother–child pairs				*Number of mother–child pairs*						
Child's gender											
Boy	41	82	55	2	3	24	35	29	32	8	1
Girl	48	77	68	3	7	31	25	30	41	7	2
Condition											
Together	42	63	58	4	—	—	—	—	—	—	—
Separate	47	96	65	1	10	55	60	59	73	15	3
Total number of pairs	89	159	123	5	10	55	60	59	73	15	3
	Average PedsQL score (SD)				*Average PedsQL score (SD)*						
Informant											
Children	37.8	37.2	37.6	37.8	64.6	71.1	73.7	74.1	74.3	72.8	67.0
	(5.1)	(4.5)	(4.7)	(2.2)	(10.2)	(10.0)	(12.2)	(10.2)	(11.0)	(7.1)	(8.9)
Mothers	41.9	41.2	40.3	39.2	75.2	77.9	77.3	75.6	75.1	76.4	68.5
	(4.4)	(5.3)	(5.4)	(6.3)	(9.1)	(8.7)	(10.4)	(9.9)	(11.5)	(8.3)	(28.2)

Note. PedsQL = pediatric health-related quality of life measure.

compute explained variance in the mixed effects model, we computed the relative decrease in the sum of the Level 1 variance (i.e., variance in G_n^p within children) and the Level 2 variance (i.e., variance in G_n^p between children) by adding the predictors (Snijders & Bosker, 1999).

Application Example. We illustrate how test developers may use the validity-index approach to estimate the age that a self-report is of equal quality as a corresponding informant-report. First, we identified the child ages for which there were enough mother–child pairs available (i.e., at least 30 per group) and excluded the mother–child pairs not corresponding to these child ages, see Table 1. This resulted in an age range of 5 to 7 years for the PedsQL$_{5\text{-}7}$ data and 8 to 11 years for the PedsQL$_{8\text{-}12}$ data.

We illustrate the validity-index approach, by (visually) comparing mean validity-index values and classification rates (i.e., the rate of response patterns classified as suspect) and corresponding 95% confidence intervals across different child-age groups. We used a small sample size adjustment for the confidence interval of the classification rate: if the number of suspect response patterns was smaller than 15 in a specific subgroup, the confidence interval was computed after adding 2 to both the original number of suspect response patterns and "normal" response patterns (Agresti & Coull, 1998). We inspected both mean validity-index values and classification rates because both provide useful information: mean values take all available information into account and do not depend on a specific cutoff value. On the other hand, differences in classification rates across groups are easier to interpret with regard to effect size. Also, group differences in classification rates have more practical value than mean differences; mean differences also include differences in validity-index values among persons with valid response patterns (e.g., differences between people with no response inconsistency and mild levels of response inconsistency) not only between persons with "normal" and suspect response patterns.

Results

Missing Values

For one mother–child pair, the PedsQL$_{5\text{-}7}$ data were excluded from all analyses due to three missing item scores in the child's response pattern (this was the only 4-year-old child in the sample). For 11 mother–child pairs, the PedsQL$_{8\text{-}12}$ data were excluded from all analyses due to three or more missing item scores in either the mother or child report. The remaining 449 (PedsQL$_{5\text{-}7}$) and 335 (PedsQL$_{8\text{-}12}$) pairs of response patterns were used for computing the endorsement rates, testing the unidimensionality and monotonicity assumptions, and to compute the cutoff values for categorizing response patterns as suspect.

For the other analyses, another 73 (PedsQL$_{5\text{-}7}$) and 60 (PedsQL$_{8\text{-}12}$) mother–child pairs were excluded because it was unclear whether the child completed the PedsQL together or separately from a parent (see the Method section). Table 1 (upper part) shows the remaining number of child–mother pairs given the child's age (in years), gender, and condition (separate vs. together). In the PedsQL$_{5\text{-}7}$ data, including 377 mother–child pairs, five children and one mother had one missing item score. In the PedsQL$_{8\text{-}12}$ data, including 274 mother–child pairs, 38 children had either one ($n = 27$) or two ($n = 11$) missing item scores. Twenty-five mothers had either one ($n = 13$) or two ($n = 8$) missing item scores.

Appropriateness Validity Indices

None of the items in the $PedsQL_{5\text{-}7}$ and $PedsQL_{8\text{-}12}$ data sets showed significant violations of the monotonicity assumption. Next, we evaluated the dimensionality of the subscales, separately in the child data, mother data, and combined child–mother data (see the appendix). In the combined data, the percentage of variance explained by the first component varied from 45 to 55 ($PedsQL_{5\text{-}7}$ data) and from 47 to 61 ($PedsQL_{8\text{-}12}$ data). Parallel analysis and the ratio between the first and second eigenvalues suggested the following for the combined mother–child data: the Emotional subscale was dominated by a single dimension for both the $PedsQL_{5\text{-}7}$ and $PedsQL_{8\text{-}12}$; the Physical and Social subscale were dominated by a single dimension for the $PedsQL_{8\text{-}12}$ but for the $PedsQL_{5\text{-}7}$ some weak multidimensionality was present; the School subscale appeared two dimensional for both the $PedsQL_{5\text{-}7}$ and $PedsQL_{8\text{-}12}$. Separate dimensionality analyses in the mother and child data suggested that lack of unidimensionality was mainly a problem in the $PedsQL_{5\text{-}7}$ child data. Taken together, the dimensionality results suggest that performance of the G_n^p and G^p indices for detecting suspect responding may be compromised due to some degree of multidimensionality, although for most subscale data sets, we found a strong first dimension.

Next, we inspected the PedsQL total score differences between mothers and children and differences across age groups. In the $PedsQL_{5\text{-}7}$ group, the mean total score was 41.0 ($SD = 5$; median = 43) for mothers and 37.5 ($SD = 4.7$; median = 38) for children. In the $PedsQL_{8\text{-}12}$ group, the mean total score for mothers was 76.3 ($SD = 10.4$; median = 78.0) and 73.0 ($SD = 10.8$; median = 74.0) for children. Table 1 (lower part) further shows that children reported consistently lower total scores (i.e., worse HRQOL) compared to mothers. Based on these mean-score differences, we chose to use G_n^p in the main analysis and used G^p to replicate the results.

Finally, we inspected the item-score distribution. As a consequence of the general population sample of healthy children, there were few responses in the response categories reflecting problematic HRQOL. In the combined $PedsQL_{5\text{-}7}$ data of mothers and children, for 11 items the number of responses in the lowest category was 15 or less. In the combined $PedsQL_{8\text{-}12}$ data, for 21 items, the number of 0-scores ranged from 2 to 9, and for eight items, the number of 1 scores ranged from 1 to 10. The skewed item-score distribution made us to adapt our data analytic strategy in two ways. First, for the $PedsQL_{8\text{-}12}$ data, we merged the two lowest categories into a single category before computing G_n^p and G^p. Second, for computing G^p and G_n^p, we calculated the average endorsement rates based on the combined data of children and mothers. Combining the data of children and mothers was expected to result in more accurate estimates of the endorsement rates, especially for the lowest response categories that were rarely endorsed. Sensitivity analyses were conducted to assess differences in results when computing G_n^p based on group-specific endorsement rates (i.e., mother's G_n^p values were based on endorsement rates computed in the mother data; children's G_n^p values were based on endorsement rates computed in the child data).

Descriptives for the Validity Index

The G_n^p index correlated negatively to the PedsQL total scores in both the $PedsQL_{5\text{-}7}$ data (Spearman's rho = −.50) and the $PedsQL_{8\text{-}12}$ data (Spearman's rho = −.32). So, for response patterns with lower total scores (i.e., lower HRQOL), the G_n^p index generally suggested more response inconsistency and therefore more problems with validity. This result can be explained by the healthy sample and resulting skewed item-score data: when providing random or inconsistent item scores, these are likely to be lower than the true item score (i.e., the score that would have been produced given accurate responding).

For mothers and children separately, Figures 1 and 2 show the G_n^p distribution in the $PedsQL_{5\text{-}7}$ data (per condition) and in the $PedsQL_{8\text{-}12}$ data. For both data sets, the differences in G_n^p distributions across subgroups seem consistent with our expectations. The center of the G_n^p distribution is closer to 0 for mothers and for children in the together condition and the relative number of outliers is larger in both subgroups where children completed the PedsQL independently. The cutoff values for classifying G_n^p values as suspect were 0.38 ($PedsQL_{5\text{-}7}$) and 0.35 ($PedsQL_{8\text{-}12}$). Sixty (6.7%) $PedsQL_{5\text{-}7}$ response patterns and 52 (7.8%) $PedsQL_{8\text{-}12}$ response patterns were classified as suspect.

Hypothesis 1: Effect of Condition and Age on Validity

The multiple linear regression model for the $PedsQL_{5\text{-}7}$ self-report data showed the expected significant negative effect of condition (separate = 0; together = 1) on G_n^p values ($b = -0.09$, $t = -4.14$) and the expected significant positive interaction effect between age and condition ($b = 0.04$, $t = 2.25$). Age had a significant negative main effect on G_n^p ($b = -0.04$, $t = -3.29$), suggesting that the validity of the response patterns improved with increasing age in the separate condition. The proportion of explained variance was 0.07, indicating a small to moderate effect size. So, consistent with our hypothesis, children aged 5 to 7 years had poorer validity-index values in the separate condition compared with the together condition, and the negative effect of the separate condition on the validity of the response patterns was smaller with increasing age.

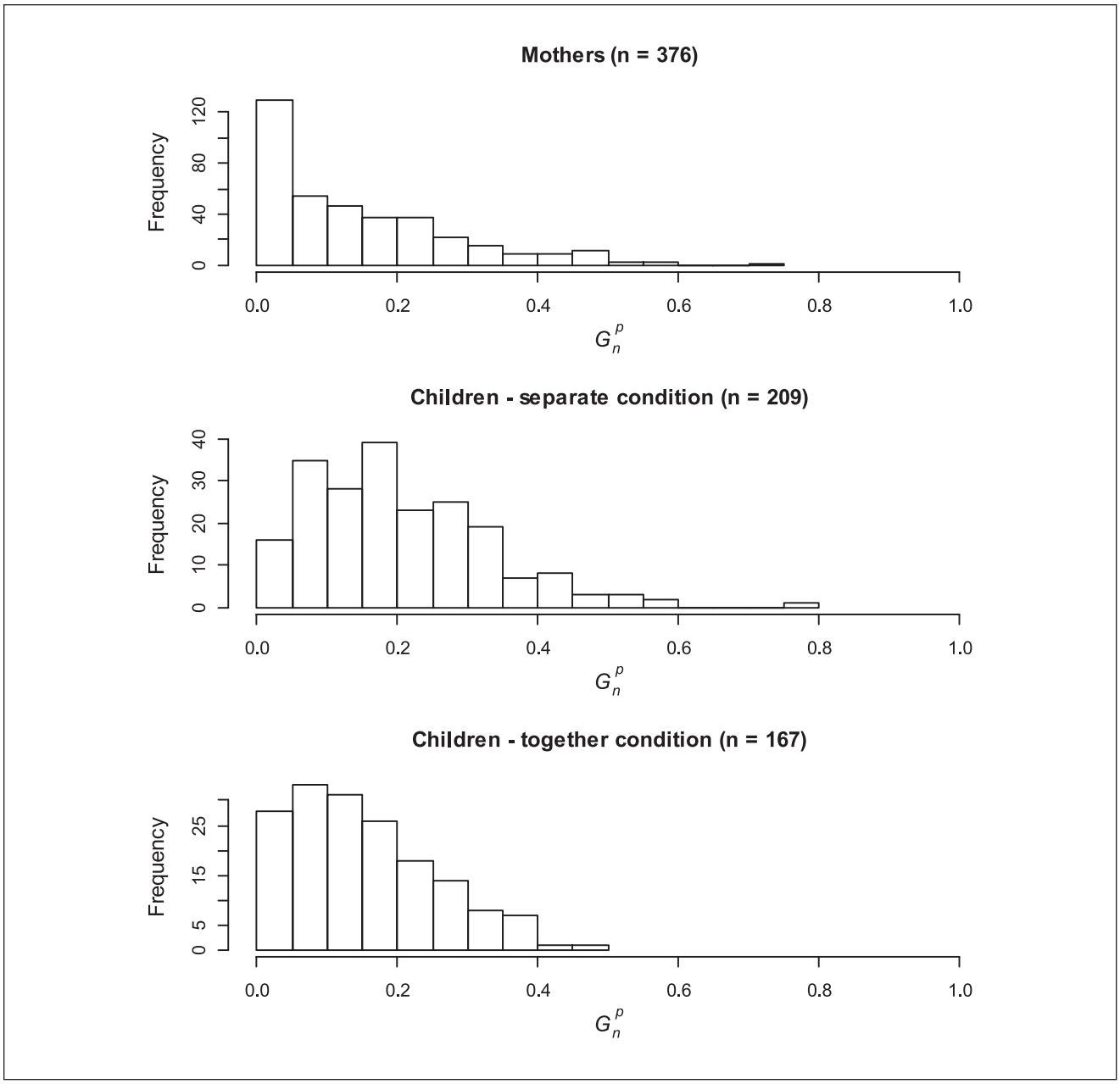

Figure 1. Distribution of G_n^p values for the $PedsQL_{5\text{-}7}$ data.
Note. PedsQL = pediatric health-related quality of life measure.

Hypothesis 2: Effect of Age and Informant on Validity

Two mixed effects models were estimated: a model for the $PedsQL_{5\text{-}7}$ data from the separate condition, and a model for the $PedsQL_{8\text{-}12}$ data (separate completion only). The estimated models showed similar results for the two scale versions: we found the expected negative effect of informant (self = 0; informant = 1) on G_n^p values ($PedsQL_{5\text{-}7}$: $b = -0.08$, $t = -4.84$; $PedsQL_{8\text{-}12}$: $b = -0.12$, $t = -6.28$) and the expected positive interaction effect for age and informant ($PedsQL_{5\text{-}7}$: $b = 0.03$, $t = 2.84$; $PedsQL_{8\text{-}12}$: $b = 0.02$, $t = 3.82$). In both models, there was a significant negative main effect of age ($PedsQL_{5\text{-}7}$: $b = -0.02$, $t = -2.25$; $PedsQL_{8\text{-}12}$: $b = -0.02$, $t = -3.68$), suggesting that the validity of the self-report improved with age. The proportion of explained variance in G_n^p values within child–mother pairs was 0.04 ($PedsQL_{5\text{-}7}$) and 0.08 ($PedsQL_{8\text{-}12}$). So, consistent with our hypotheses, for both data sets we found that children's response patterns had poorer validity-index values compared to their mother's informant reports and that this negative effect of self-report (vs. informant-report) on validity was smaller for older children.

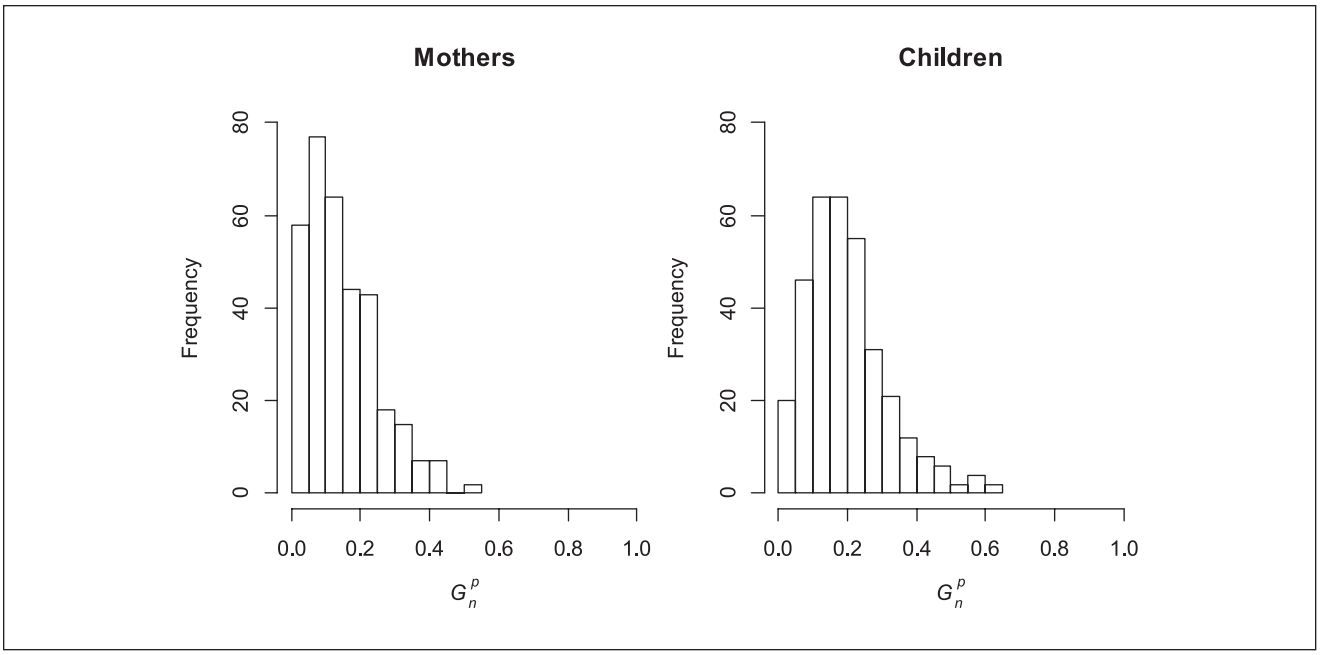

Figure 2. Distribution of G_n^p values for the $PedsQL_{8\text{-}12}$ data ($n = 335$ mother–child pairs).
Note. PedsQL = pediatric health-related quality of life measure.

Application Example

Child–mother pairs in a child-age category with 30 or less pairs were excluded from the application example, resulting in a minimum of 42 pairs in each condition subgroup (see Table 1). Data of 208 ($PedsQL_{5\text{-}7}$; separate condition only) and 246 ($PedsQL_{8\text{-}12}$) child–mother pairs were left for the analysis. Two separate sets of analyses were conducted as the data came from two different PedsQL versions. Figures 3 and 4 (left panels) show the average G_n^p values and classification rates for children and mothers as a function of the child's age group, for each PedsQL version, respectively. Confidence intervals around the classification rates show that the rates are estimated with a considerable margin of error, so differences in classification rates are mainly used descriptively and as an indication of effect size.

For the $PedsQL_{5\text{-}7}$, the average G_n^p values suggest that from 7 years on, there is no difference in validity between mother reports and child reports (Figure 3). The classification rates further suggest that the difference in validity is only meaningful for the 5-year-old. The classification rates show a substantially higher rate of self-reports classified as severely inconsistent in the 5-year-old subgroup ($n = 9$; 19%) compared with the rate of informant-reports classified a severely inconsistent ($n = 1$; 2%). For the 6- and 7-year-old, there is no difference in classification rates between self-reports and informant-reports. These results suggest that from 6 years on, children can provide interview-based $PedsQL_{5\text{-}7}$ self-report data, without the help of their parents, of similar quality as informant-report data.

For the $PedsQL_{8\text{-}12}$, the average G_n^p values suggest that from age 9 on, there is no significant difference in validity between the self-report and the informant-report (Figure 4). The classification rates further show that for none of the age groups there are large differences in classification rates. The number of severely inconsistent self-reports in the 8-year-old group ($n = 8$; 14.5%) and 9-year-old group ($n = 7$; 11.7%) is somewhat larger than the number of severely inconsistent informant-reports for 8-year-old ($n = 3$; 5.0%) and 9-year-old ($n = 3$; 5.4%). For ages 10 and 11 years, there are no differences in classification rates between mother reports and child reports. So, for the $PedsQL_{8\text{-}12}$, results suggest that the 8-year-old tend to respond more inconsistently compared with parents but the effect of the response inconsistencies on test score validity may be negligible given the modest differences in classification rates. The results are in accordance with previous research showing that 8- and 9-year-old children show more extreme response styles than adults or older children (Chambers & Johnston, 2002; Davis et al., 2007) but children aged 8 years and older still provide sufficiently valid and reliable $PedsQL_{8\text{-}12}$ self-report data (e.g., Varni et al., 2007a, 2007c).

Sensitivity Analyses

Hypothesis Tests. We repeated the regression analyses regarding Hypotheses 1 and 2 using two alternative dependent variables: the G^p index and the group-specific G_n^p index. For the G^p index, the analyses were slightly

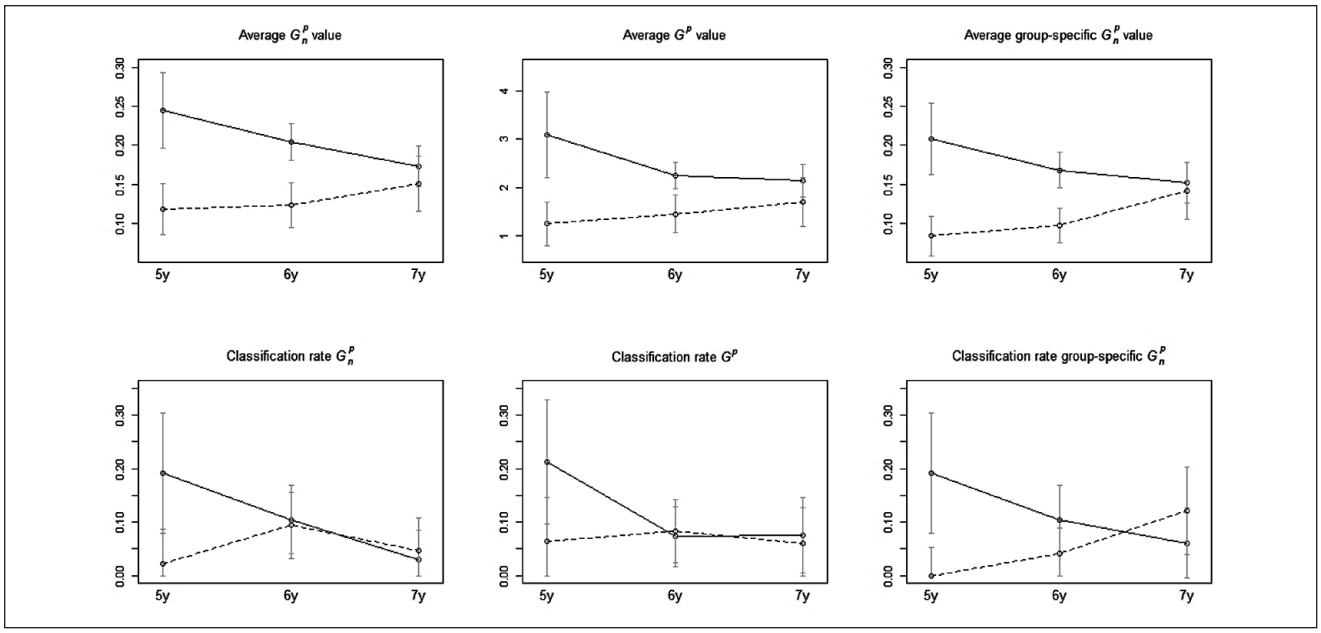

Figure 3. Validity-index means and classification rates per age group, including 95% confidence intervals, for $PedsQL_{5\text{-}7}$ data of children (solid line) and mothers (dashed line) in the "separate" condition.
Note. PedsQL = pediatric health-related quality of life measure.

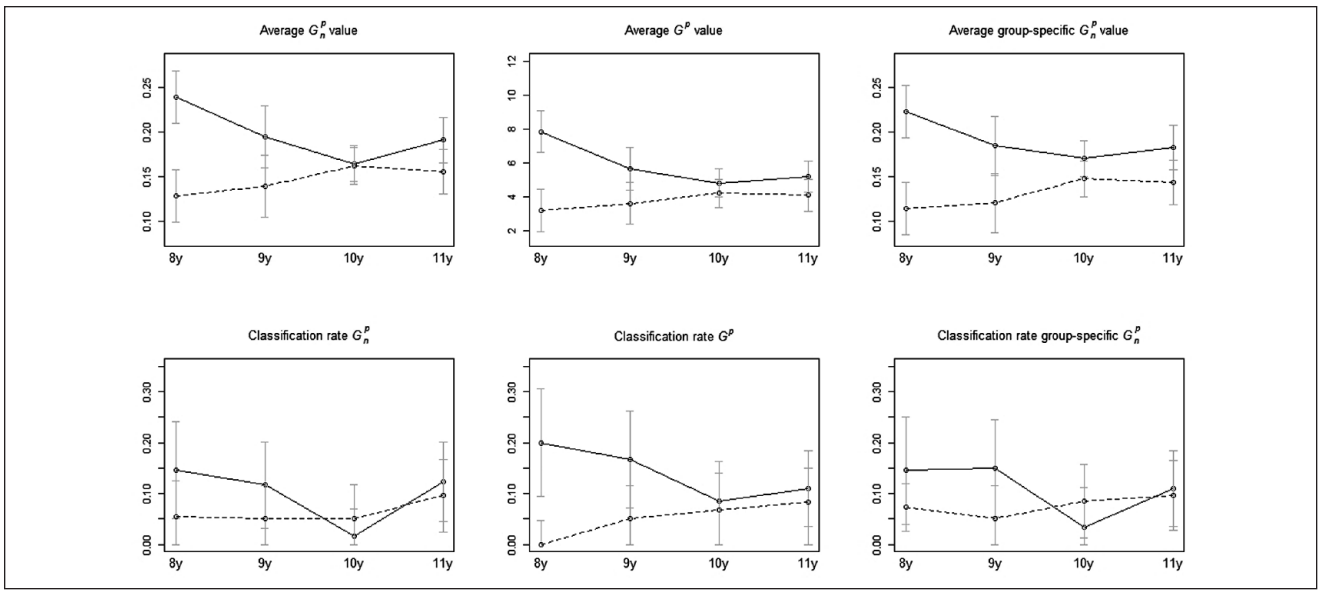

Figure 4. Validity-index means and classification rates per age group, including 95% confidence intervals, for the $PedsQL_{8\text{-}12}$ data of children (solid line) and mothers (dashed line).
Note. PedsQL = pediatric health-related quality of life measure.

different: we now controlled for the PedsQL total scores in the analyses, as G^p depends on the total score. G_n^p and G^p were highly correlated in the $PedsQL_{5\text{-}7}$ data (Spearman's rho = .85) and the $PedsQL_{8\text{-}12}$ data (Spearman's rho = .75). G_n^p and the group-specific G_n^p were also correlated highly both in the $PedsQL_{5\text{-}7}$ data (Spearman's rho = .89) and the $PedsQL_{8\text{-}12}$ data (Spearman's rho = .88).

For the G^p index, conclusions regarding both hypotheses equaled those of the main analysis. For the group-specific G_n^p index, we could replicate the conclusions regarding Hypothesis 2, but we could not fully replicate the conclusions regarding Hypothesis 1: we found the significant negative effect of condition but not the expected interaction effect of age and condition. Figure 5 (right panel)

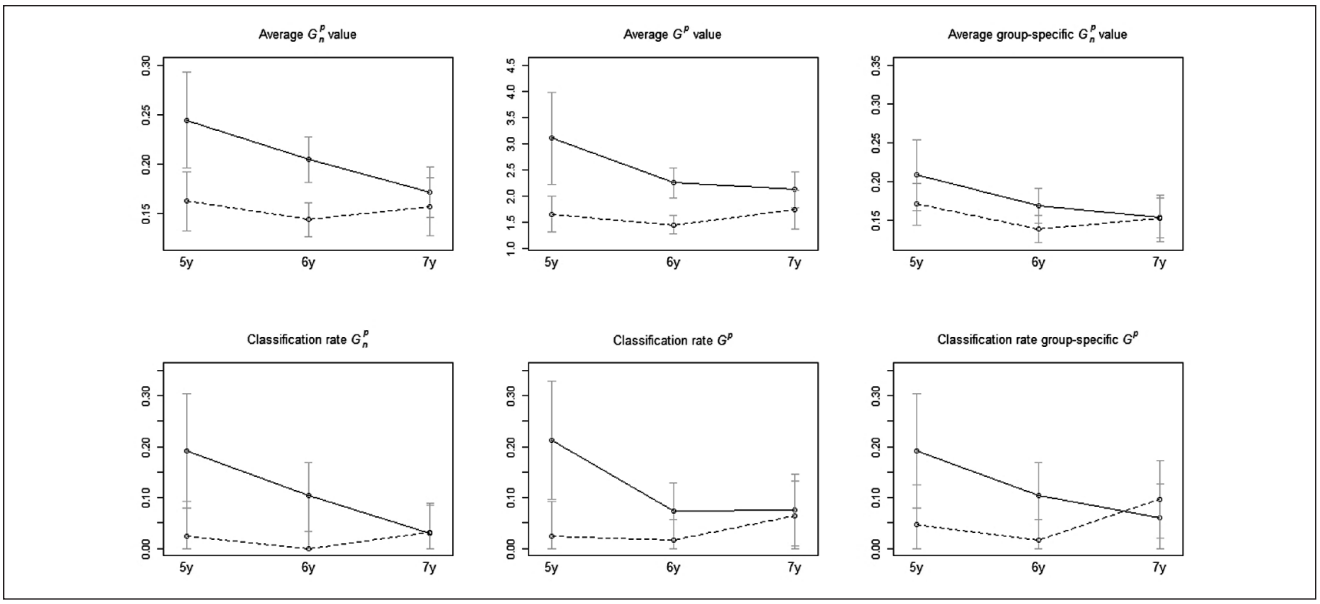

Figure 5. Validity-index means and classification rates per age group, including 95% confidence intervals, for PedsQL$_{5\text{-}7}$ data of children in the separate condition (solid line) and together condition (dashed line).
Note. PedsQL = pediatric health-related quality of life measure.

shows that the pattern of means is generally as expected, but the effects are smaller compared with those of the main analysis using G_n^p.

Application Example. Figures 3 and 4 show that the pattern of means and classification rates for both the G^p index and the group-specific G_n^p index are similar to that of G_n^p, and the same conclusions would be drawn. The most notable difference in results is the larger difference in classification rates based on G^p between self-reports (0.20) and mother reports (0.00) in the PedsQL$_{8\text{-}12}$ data for the 8-year-old.

Discussion

Various researchers have stressed the importance of studying whether and at what age young children provide valid self-reports (e.g., Landgraf et al., 2018; U.S. Food and Drug Administration, 2009). We proposed the validity-index approach to estimate the minimum age for a child to produce self-report data with similar quality as that of an adult informant. Data of a popular pediatric HRQOL measure, the PedsQL, were used to evaluate and illustrate the approach, but the approach is potentially suitable for a wide range of measures, such as personality inventories and psychopathology scales.

Preliminary analyses showed that indices to detect repetitive responding or response biases were not suitable for application to our PedsQL data, and we therefore continued our analyses using a person-fit index to detect inconsistent responding. Application of the validity-index approach to the PedsQL scales, using person-fit statistic G_n^p, confirmed our hypotheses regarding the effects of the child's age, the informant, and administration method on validity. These results suggest that person-fit statistics such as G_n^p can discriminate between groups expected to differ in the validity of their responses, and are useful indices for application in the validity-index approach. In general, the pattern of subgroup differences in validity was also consistent using different independent variables (e.g., classification rates vs. mean validity-index values and G_n^p vs. G^p). So, our approach seems to be robust with regard to specific analytic choices.

In future research, the validity-index approach can be easily implemented in standard validation research; in child research, self-report and informant-data are commonly collected during the process of questionnaire validation, and substantial sample sizes for different age groups are also routinely collected for establishing reference values. As explained in the Method section, we recommend that subgroup differences in both mean validity-index values and classification rates are analyzed when using the validity-index approach. However, an important drawback of analyzing classification rates—also experienced in our study—is that when the sample size is not very large, the estimated differences in classification rates across groups can be unreliable due to sampling fluctuation.

Our illustrative application provides recommendations on the minimum ages for self-reporting on the two PedsQL versions. The two most important results are the following. First, a substantial part of 5-year-old could not

independently provide PedsQL$_{5\text{-}7}$ self-report data with sufficient psychometric quality. Second, we showed that the minimum age for providing a self-report of sufficient quality may depend on scale properties, such as the response scale; our results indicate that 7-year-old children provided self-report data on the PedsQL$_{5\text{-}7}$ (3-point scale) with similar quality as informant data from adults, but on the PedsQL$_{8\text{-}12}$ (5-point scale) 8-year-old showed more inconsistent responding compared with adult informants. It should be noted that the current results apply to the sample and scales used, and the conclusions with reference to the minimally required age for self-reporting may not generalize to other populations and questionnaires.

Future applications of the approach may also use other independent variables in a modified validity-index approach: the child's age in months (instead of age in years), reading-level, or intelligence test scores. The independent variables explaining most variance in validity-index values can be expected to be most informative to determine which child properties are needed for a self-report of sufficient quality. Another interesting topic for future research is to extend the validity-index approach to include data of multiple adult informants, such as both parents.

Strengths and Limitations

An important strength of the current study is that we could illustrate the validity-index method using PedsQL data; identifying the minimum age for a self-report is a particularly important topic in HRQOL research (e.g., Eiser et al., 2000). Future research in this assessment context may benefit from our application example providing preliminary results regarding the minimum age for the PedsQL$_{5\text{-}7}$ and the PedsQL$_{8\text{-}12}$. The main limitations of the current study are, however, also related to the specific data set. A first important drawback of the PedsQL data set was that only indices for detecting inconsistent responding were suitable for the data, while indices of repetitive responding and response bias were too strongly confounded with the respondents' underlying trait value due to specific data and scale properties. As a consequence, we could not illustrate the potential of the validity-index approach for studying different types of children's aberrant response behavior. The conclusions from the application example should be evaluated within the context of this limitation. The children (and mothers) may have used other types of invalid response styles that were unrelated or even opposite to response inconsistency (e.g., repetitive responding). A second limitation of the data set is the variable participant rate; the effect of selection bias should therefore not be underestimated. Hence, the specific results concerning the PedsQL$_{5\text{-}7}$ and the PedsQL$_{8\text{-}12}$ should be replicated and extended in future studies to provide final recommendations to practitioners.

We applied one of the most powerful statistics available for detecting inconsistent responding (e.g., Niessen et al., 2016). However, also for detecting inconsistent responding, the PedsQL does not have optimal properties and this may have resulted in relatively low detection rates. First, the unidimensional PedsQL subscales are short (five to eight items) and the total number of items is small (23); this small number of items results in low power to detect invalid responding (Conijn et al., 2014; Emons, 2008). Second, in the PedsQL, items are presented per subscale instead of in a mixed order. Third, all items are formulated in the same direction. The two latter properties also increase the possibility that invalid response patterns go undetected (i.e., Type II errors) because respondents may produce consistent patterns of item scores without comprehending or reading items well. Given these limitations, it is actually remarkable that our method worked rather well for the PedsQL when using the Guttman person-fit statistics.

We therefore believe that the validity-index approach has shown good potential and should also be evaluated using other types of questionnaires, such as personality inventories assessing the Big Five personality traits (e.g., the 144-item Hierarchical Personality Inventory for Children; Mervielde & De Fruyt, 1999). More validity indices can then be added to the approach, providing a more complete image of the children's development in generating self-report data. If a questionnaire includes more mixed content (subscales measuring unrelated constructs) or both positively and negatively worded items measuring the same construct, the Long string index and response bias indices can be expected to be useful as well (DeSimone et al., 2015; Van Herk et al., 2004). And when questionnaires include many items (i.e., 50 or more), different types of invalid responding can be measured more reliably than using the 23-item PedsQL scale.

Conclusion

Our results suggest that the validity-index approach can be used to study at what age children provide valid self-reports on a given measure. The approach appeared to work remarkably well using a relatively short instrument. However, the PedsQL data were not optimal for showing the full potential of our suggested approach. Future research should evaluate the validity-index approach using questionnaires with more items and more mixed content, and use multiple validity indices addressing different types of invalid responding.

Appendix

Results of the Dimensionality Analyses for the $PedsQL_{5-7}$ Data (n = 449 Child–Mother Pairs) and $PedsQL_{8-12}$ Data (n = 335 Child–Mother Pairs).

		$PedsQL_{5-7}$				$PedsQL_{8-12}$			
Subscale	Subsample	Percentage variance explained first component	Percentage variance explained second component	Number of components based on parallel analysis	λ_2/λ_1	Percentage variance explained first component	Percentage variance explained second component	Number of components based on parallel analysis	λ_2/λ_1
Physical functioning	Mothers	56	12	1	0.21	66	11	1	0.16
	Children	36	17	2	0.46	45	12	1	0.26
	Combined	45	14	2	0.31	55	11	1	0.20
Emotional functioning	Mothers	59	15	1	0.26	62	14	1	0.23
	Children	42	18	1	0.46	48	18	1	0.37
	Combined	50	16	1	0.33	53	16	1	0.29
Social functioning	Mothers	66	13	1	0.19	63	19	1	0.29
	Children	42	19	1	0.45	54	14	1	0.27
	Combined	55	16	2	0.29	61	16	1	0.26
School functioning	Mothers	54	25	2	0.47	50	24	2	0.48
	Children	41	25	2	0.62	44	25	2	0.58
	Combined	48	27	2	0.56	47	28	2	0.58

Note. PedsQL = pediatric health-related quality of life measure.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Aaronson, N. K., Meyerowitz, B. E., Bard, M., Bloom, J. R., Fawzy, F. I., Feldstein, M., . . . Ware, J. E. (1991). Quality of life research in oncology: Past achievements and future priorities. *Cancer*, *67*, 839-843.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician*, *52*, 119-126.

Bastiaansen, D., Koot, H. M., Ferdinand, R. F., & Verhulst, F. C. (2004). Quality of life in children with psychiatric disorders: Self-, parent, and clinician report. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*, 221-230.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48.

Burns, M. K. (2002). Self-report objective measures of personality for children: A review of psychometric properties for RQC. *Psychology in the Schools*, *39*, 221-234.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29.

Chambers, C. T., & Johnston, C. (2002). Developmental differences in children's use of rating scales. *Journal of Pediatric Psychology*, *27*, 27-36.

Cole, D. A., Goodman, S. H., Garber, J., Cullum, K. A., Cho, S.-J., Rights, J. D., Felton, J. W., . . . Simon, H. F. M. (2018). Validating parent and child forms of the Parent Perception Inventory. *Psychological Assessment*, *30*, 1065-1081. doi:10.1037/pas0000552

Conijn, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic $\boldsymbol{l}_z$-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, *38*, 122-136.

Conijn, J. M., Franz, G., Emons, W. H. M., De Beurs, E., & Carlier, I. V. E (2019). The assessment and impact of careless responding in routine outcome monitoring within mental health care. *Multivariate Behavioral Research*, *54*, 593-611.

Conijn, J. M., van der Ark, A., & Spinhoven, P. (2020). Satisficing in mental health care patients: The effect of cognitive symptoms on self-report data quality. *Assessment*, *27*, 178-193. doi:10.1177/1073191117714557

Davis, E., Nicolas, C., Waters, E., Cook, K., Gibbs, L., Gosch, A., & Ravens-Sieberer, U. (2007). Parent-proxy and child self-reported health-related quality of life: Using qualitative methods to explain the discordance. *Quality of Life Research*, *16*, 863-871.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483-509.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*, 171-181.

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior—theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*, 558-574.

Eddy, L., Khastou, L., Cook, K. F., & Amtmann, D. (2011). Item selection in self-report measures for children and adolescents with disabilities: Lessons from cognitive interviews. *Journal of Pediatric Nursing*, *26*, 559-565.

Eiser, C., Mohay, H., & Morse, R. (2000). The measurement of quality of life in young children. *Child: Care, Health and Development*, *26*, 401-414.

Eiser, C., & Varni, J. W. (2013). Health-related quality of life and symptom reporting: Similarities and differences between children and their parents. *European Journal of Pediatrics*, *172*, 1299-1304.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224-247.

Ferreira, P. L., Baltazar, C. F., Cavalheiro, L., Cabri, J., & Gonçalves, R. S. (2014). Reliability and validity of PedsQL for Portuguese children aged 5-7 and 8-12 years. *Health and Quality of Life Outcomes*, *12*, 122. doi:10.1186/s12955-014-0122-3

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99-114.

Huebner, E. S. (1991). Initial development of the Students' Life Satisfaction Scale. *School Psychology International*, *12*, 231-240.

Jansen, M., Bodden, D. H., Muris, P., van Doorn, M., & Granic, I. (2017). Measuring anxiety in children: The importance of separate mother and father reports. *Child & Youth Care Forum*, *46*, 643-659.

Jozefiak, T. (2014). Can we trust parents' report about their children's well-being? In A. Ben-Arieh, F. Casas, I. Frønes & J. E. Korbin (Eds.), *Handbook of child well-being: Theories, methods and policies in global perspective* (pp. 577-578). Dortmund, Netherlands: Springer.

Kobayashi, K., & Kamibeppu, K. (2010). Measuring quality of life in Japanese children: Development of the Japanese version of PedsQL. *Pediatrics International*, *52*, 80-88.

Landgraf, J. M., van Grieken, A., & Raat, H. (2018). Giving voice to the child perspective: Psychometrics and relative precision findings for the Child Health Questionnaire self-report short form (CHQ-CF45). *Quality of Life Research*, *27*, 2165-2176.

Limbers, C. A., Newman, D. A., & Varni, J. W. (2008). Factorial invariance of child self-report across age subgroups: A confirmatory factor analysis of ages 5 to 16 years utilizing the PedsQL 4.0 Generic Core Scales. *Value in Health*, *11*, 659-668.

Matza, L. S., Swensed, A. R., Flood, E. M., Secnik, K., & Leidy, N. K. (2004). Assessment of health-related quality of life in children: Review of conceptual, methodological, and regulatory issues. *Value in Health*, *7*, 79-92.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437-455.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*, 311-314.

Mervielde, I., & De Fruyt, F. (1999). Construction of the Hierarchical Personality Inventory for Children (HiPIC). In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality psychology in Europe: Proceedings of the eight European Conference on Personality Psychology* (pp. 107-127). Tilburg, Netherlands: Tilburg University Press.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1-11.

Peer, E., & Gamliel, E. (2011). Too reliable to be true? Response bias as a potential source of inflation in paper-and-pencil questionnaire reliability. *Practical Assessment, Research & Evaluation*, *16*, 1-8.

Pereira, A. I., Muris, P., Barros, L., Goes, R., Marques, T., & Russo, V. (2015). Agreement and discrepancy between mother and child in the evaluation of children's anxiety symptoms and anxiety life interference. *European Child & Adolescent Psychiatry*, *24*, 327-337.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raiche, G. (2010). *nFactors: An R package for parallel analysis and non graphical solutions to the Cattell scree test*. Retrieved from https://cran.r-project.org/web/packages/nFactors/nFactors.pdf

Reeve, B. B., McFatrich, M., Pinheiro, L. C., Weaver, M. S., Sung, L., Withycombe, J. S., . . . Tomlinson, D. (2017). Eliciting the child's voice in adverse event reporting in oncology trials: Cognitive interview findings from the Pediatric Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events initiative. *Pediatric Blood & Cancer*, *64*, e26261.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3-25.

Seo, D. G., & Weiss, D. J. (2013). $\boldsymbol{l}_z$ Person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement*, *73*, 994-1016.

Shelton, K. K., Frick, P. J., & Wootton, J. (1996). Assessment of parenting practices in families of elementary school-age children. *Journal of Clinical Child Psychology*, *25*, 317-329.

Sinharay, S. (2017). How to compare parametric and nonparametric person-fit statistics using real data. *Journal of Educational Measurement*, *54*, 420-439.

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*, 443-461.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Solans, M., Pane, S., Estrada, M., Serra-Sutton, V., Berra, S., Herdman, M., . . . Rajmil, L. (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.

Taber, S. M. (2010). The veridicality of children's reports of parenting: A review of factors contributing to parent–child discrepancies. *Clinical Psychology Review*, *30*, 999-1010.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. N. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*, i05. doi:10.18637/jss.v074.i05

Thompson, H. L., Reville, M.-C., Price, A., Reynolds, L., Rodgers, L., & Ford, T. (2014). The Quality of Life Scale for Children (QoL-C). *Journal of Children's Services*, *9*, 4-17.

Tsze, D. S., von Baeyer, C. L., Bulloch, B., & Dayan, P. S. (2013). Validation of self-report pain scales in children. *Pediatrics*, *132*, e971-e979.

U.S. Food and Drug Administration. (2009). *Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims*. Retrieved from https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf

van der Ark, A. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, *20*, i11. doi:10.18637/jss.v020.i11

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*, 346-360.

Varni, J. W., Limbers, C. A., & Burwinkle, T. M. (2007a). How young can children reliably and validly self-report their health-related quality of life? An analysis of 8,591 children across age subgroups with the PedsQL™ 4.0 Generic Core Scales. *Health and Quality of Life Outcomes*, *5*, 1. doi:10.1186/1477-7525-5-1

Varni, J. W., Limbers, C. A., & Burwinkle, T. M. (2007b). Literature review: Health-related quality of life measurement in pediatric oncology: Hearing the voices of the children. *Journal of Pediatric Psychology*, *32*, 1151-1163.

Varni, J. W., Limbers, C. A., & Burwinkle, T. M. (2007c). Parent proxy-report of their children's health-related quality of life: An analysis of 13,878 parents' reliability and validity across age subgroups using the PedsQL™ 4.0 Generic Core Scales. *Health and Quality of Life Outcomes*, *5*, 2. doi:10.1186/1477-7525-5-2

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*, 800-812.

Varni, J. W., Seid, M., & Rode, C. A. (1999). The PedsQL™: Measurement model for the Pediatric Quality of Life Inventory. *Medical Care*, *37*, 126-139.

Wallander, J. L., & Koot, H. M. (2016). Quality of life in children: A critical examination of concepts, approaches, issues, and future directions. *Clinical Psychology Review*, *45*, 131-143.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*, 102-117.

Woolley, M. E., Bowen, G. L., & Bowen, N. K. (2004). Cognitive pretesting and the developmental validity of child self-report instruments: Theory and applications. *Research on Social Work Practice*, *14*, 191-200.

World Health Organization. (1947). *Constitution of the World Health Organization*. Retrieved from https://www.who.int/governance/eb/who_constitution_en.pdf

Yi-Frazier, J. P., Hilliard, M. E., Fino, N. F., Naughton, M. J., Liese, A. D., Hockett, C. W., . . . Lawrence, J. M. (2016). Whose quality of life is it anyway? Discrepancies between youth and parent health-related quality of life ratings in type 1 and type 2 diabetes. *Quality of Life Research*, 25, 1113-1121. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4936832/

Zapolski, T. C., & Smith, G. T. (2013). Comparison of parent versus child-report of child impulsivity traits and prediction of outcome variables. *Journal of Psychopathology and Behavioral Assessment*, *35*, 301-313.