European Child & Adolescent Psychiatry (2020) 29:123–136
https://doi.org/10.1007/s00787-019-01337-1

ORIGINAL CONTRIBUTION

Age-specific norms and validation of the German SDQ parent version based on a nationally representative sample (KiGGS)

Silke Janitza[1] · **Kathrin Klipker**[1,2] · **Heike Hölling**[1]

Received: 3 December 2018 / Accepted: 11 April 2019 / Published online: 23 April 2019

Abstract
The Strengths and Difficulties Questionnaire (SDQ) is the most widely used mental health screening instrument for children and adolescents. It is a short questionnaire including 25 items that can be answered by parents, teachers or children. There are two studies which report norms for the German SDQ parent version. They do not include children younger than 6 years. Moreover, whether the German SDQ parent version is measurement invariant across age has not yet been investigated. The absence of measurement invariance across age would support the use of age-specific norms that are not yet available for the German SDQ parent version. We used data of the German Health Interview and Examination Survey for Children and Adolescents (KiGGS), a nationally representative survey including 14,835 children aged 3–17 years, to assess measurement invariance of the German SDQ parent version across the full age range. Multi-group confirmatory factor analysis revealed that the hyperactivity and emotional symptoms subscales are not comparable between children of different ages. This supports the use of age-specific norms for these two subscales and for the total SDQ. We used methods of centile estimation to smoothly model the centiles of the SDQ total score and the subscale scores in dependence on age. These age-specific centiles reflect the developmental course of SDQ problems in children (including preschoolers) and adolescents living in Germany. They can be used to identify children and adolescents with abnormal behaviour, while accounting for the developmental course of emotional and behaviour problems.

Keywords Strength and Difficulties Questionnaire (SDQ) · Parent report · Sex- and age-specific norms · Reference curves · National norms · Screening instrument

Introduction

The identification of mental health problems in children and adolescents poses several challenges on mental health and research professionals. One important criterion of mental health problems, amongst others, is that they are characterised by a deviation from an appropriate reference group. An appropriate reference group might be a group of persons with similar demographic characteristics, such as age, sex and cultural background (e.g., [1, 2]). It is important that screening instruments are validated in nationally representative samples in different age groups and for both boys and girls before they are used for identifying mental health problems.

The Strengths and Difficulties Questionnaire (SDQ) is one of the most widely used mental health screening instruments for children and adolescents, and has been translated into over 60 languages [3, 4]. It comprises the five subscales emotional symptoms, peer problems, conduct problems, hyperactivity/inattention and prosocial behaviour. The SDQ was originally developed for children and adolescents

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00787-019-01337-1) contains supplementary material, which is available to authorized users.

✉ Silke Janitza
silke.janitza@gmail.com

[1] Department of Epidemiology and Health Monitoring, Robert Koch Institute, General-Pape-Straße 62-64, 12101 Berlin, Germany

[2] Vivantes Clinic of Child and Adolescent Psychiatry and Psychotherapy, Landsberger Allee 49, 10249 Berlin, Germany

of ages 4–17 years. An early-year version of the SDQ for children of age 2–4 years was developed in 2014.[1] The evaluation can be done by parents and teachers of children and adolescents aged 2–17 years (parent/teacher version) or by the children themselves if they are 11 years or older (self-report version). The SDQ parent version (SDQ-P) is the most widely used version [5].

Epidemiological studies showed that psychopathological abnormalities are prevalent in about 10–20% of children and adolescents [6]. Many countries have national norms that were derived from nationally representative samples (for a review, see [5]). Some authors accounted for potential sex and age differences. Based on the norms, cutoff scores defining a normal range for the SDQ total sum score can be derived for the screening of mental health problems. Due to prevalence estimates between 10% and 20%, the 80th and 90th centiles of nationally representative samples have been frequently used as cutoff scores for defining clinically relevant behaviour. More recently, Goodman et al. proposed using the 80th, 90th and 95th centiles as cutoffs for defining the scores as 'close to average' (< 80th centile), 'slightly raised' (80th–90th centile), 'high' (90th–95th centile) and 'very high' (> 95th centile) [7]. Such information could be used to interpret the severity of abnormality.

The German SDQ-P has been tested and validated in five studies [8–13]. Two of the studies were based on clinical samples comprising children aged 6–18 years with ADHD [8] and children aged 5–17 years with any psychiatric diagnosis [9]. The three remaining studies examined community samples that were considered nationally representative. Norms of the German SDQ-P stem from Woerner et al. and are based on parents' reports of approximately 1000 children aged 6–16 years [12, 13]. These norms have been confirmed by Rothenberger et al. in a sample of approximately 2500 parents of 7–16-year-old children [11]. However, there are no studies examining preschoolers, limiting the generalisability of the results to children of 6 years or older. The availability of German norms for preschoolers is inevitable because preschoolers are not yet able to report on their mental health and parents' or carers' reports are the main source of information (for a review, see [14]).

Emotional or behaviour problems might express differently depending on a child's age [1, 2]. The behaviour of an 11-year-old might be of clinical relevance, although the same behaviour might be normal at the age of, say, 3 years. Therefore, because the SDQ-P can be applied to the wide age range from 2 to 17 years, developmental differences in the phenotype of emotional or behaviour problems might arise. This would prohibit any comparisons of (subscale) scores for children of different ages. Further, it raises the concern that younger children are reported systematically more or less frequently of showing abnormal behaviour than older children, if norms of older children are used (e.g., [15]). Examination of measurement invariance of the SDQ across age is, therefore, of crucial importance and if the German SDQ-P is not measurement invariant across age, age-specific norms are required. If measurement invariance can be shown in contrast, normal ranges (usually defined by centiles of a reference population) are constant across different ages and there is no need for age-specific reference values.

Woerner et al. reported that for children and adolescents aged 6–16 years scale scores are unrelated to age, except for the hyperactivity/inattention subscale for which younger children had slightly higher scores than older children and adolescents [12, 13]. The authors concluded that the differences in the German SDQ-P across age groups have shown to be negligible. Rothenberger et al. obtained similar results for children and adolescents aged 7–16 years [11]. Again, younger children had significantly higher hyperactivity scores than older children and adolescents. In addition, younger children had slightly higher SDQ total scores and lower prosocial behaviour scores than older children. This suggests that developmental processes should be taken into account at least for some subscales. However, their investigations are limited to children of age 6–16 years and do not reveal the developmental course of SDQ problems from preschool to school-aged children. Within their preschool sample, Klein et al. obtained no age differences between 3- and 5-year-olds [10]. In addition, they compared SDQ scale means of their 3–5-year-olds with the scale means of the German representative sample of 6–16-year-olds of Woerner et al. [12, 13]. Despite the lack of representativeness and comparability, Klein et al. attribute differences between the samples to age rather than other factors, leading to conclusions that are in conflict with the available literature [10]: The authors concluded that for 3–5-year-olds prosocial behaviour and hyperactivity scores were higher and peer problem scores were lower than in 6–16-year-olds. Hölling et al., in contrast, showed that 3–6-year-olds and 14–17-year-olds had lower problem behaviour than 7–13-year-olds [16].

In all the studies, either Spearman correlation analysis or classical statistical tests and descriptive analysis were performed [12, 13]. While the former tests for the existence of any monotone relationship between the SDQ total score and age, the latter assesses whether the central tendency (like mean or median) is different across age groups. Although addressing interesting questions, neither type of analysis answers the question of whether the SDQ operates the same way for all age groups (i.e., if the SDQ is measurement invariant).

[1] The original SDQ differs from this early-year version in the wording of three items (two items of the behaviour problem subscale and one item of the hyperactivity subscale; see http://www.sdqinfo.org).

We used data of the German Health Interview and Examination Survey for Children and Adolescents (KiGGS) [17], to address the following goals: (1) to replicate the original scale structure of the German SDQ-P in the whole sample, to validate the German SDQ-P across the whole age range, including preschool children, (2) to assess whether the German SDQ-P is measurement invariant across the full age range, and (3) to provide norms that are age-specific if measurement invariance cannot be shown and not age-specific if measurement invariance can be shown. All analyses are performed separately for boys and girls because of substantial sex differences in problem behaviour [3, 16, 18].

Methods

Study design and sample

The German Health Interview and Examination Survey for Children and Adolescents (KiGGS) is a nationally representative cross-sectional health interview and examination survey for children and adolescents that took place in Germany from May 2003 to May 2006 [17]. The KiGGS Study was approved by the Charité/Universitätsmedizin Berlin Ethics Committee and the Federal Office for the Protection of Data, and was conducted according to the Declaration of Helsinki. Details on the objective, study design and sampling strategy were described elsewhere [19–21].

Participants were sampled based on a complex two-stage sampling procedure, with 167 sample points from an inventory of German communities stratified according to the BIK classification system. Data on sociodemographic characteristics as well as parameters related to physical, psychological and social health were obtained from 17,640 children and adolescents for the entire age range from 0 to 17 years. Among those, 14,835 children were 3 years or older.

Instruments

The Strengths and Difficulties Questionnaire (SDQ) is a widely used screening instrument for emotional and behaviour problems, and also contains a subscale on prosocial behaviour. It consists of 25 items positively or negatively worded, thereby assessing both strengths and difficulties. In the SDQ-P, the items are rated by parents as untrue (corresponding to a score of 0), somewhat true (score of 1) or certainly true (score of 2). They can be grouped into four problem subscales and one competence subscale, each comprising five items with sum scores ranging from 0 to 10. The four problem subscales assess conduct problems, hyperactivity/inattention, emotional symptoms and peer problems. The competence subscale assesses prosocial behaviour. The total difficulties score is composed of the scores of the problem subscales only and thus ranges from 0 to 40, with larger scores suggesting greater problem behaviour.

Statistical analysis

Derivation of SDQ subscale scores in the presence of missing data

If the parent answered three or more items of a given subscale, the respective subscale score was derived. Current practice consists in imputing the scores of items that were not answered using the mean score of the subscale, if no more than two answers per subscale are missing. For example, if a parent rates the first three items of a subscale with 1, 0 and 1, but does not rate the other two items of the subscale, the scores for these two items are imputed by the value 0.667 ($=[1+0+1]/3$). This gives a subscale score of 3.333 ($=1+0+1+0.667+0.667$). If the items of all other subscales are answered, one obtains a real-valued SDQ total score.

Note that the derivation is slightly different from the original algorithm provided at http://www.sdqinfo.org. Real-valued scores are rounded to the nearest integer (see http://www.sdqinfo.org). Rounding the score to the nearest integer, however, introduces a bias. In the example above, the value 3.333 is rounded to 3, which leads to an underestimation of the child's problem score. Moreover, according to the implementations provided at http://www.sdqinfo.org, the SDQ total score is computed from the rounded subscale scores. This might be a problem. For example, if 3.333 is the subscale score of all four subscales, the derived SDQ total score is 12 ($=3+3+3+3$), although the value 13 ($\approx 3.333+3.333+3.333+3.333$) would be more appropriate. This introduces an additional bias. For this reason, we did not round any values to the nearest integer.

Confirmatory factor analysis (CFA) for replicating the five-factor structure of the SDQ-P

CFA was used to evaluate the five-factor structure of the SDQ, based on the sample population for which all 25 items of the SDQ were answered. The analysis was performed both for the total sample population and separately for boys and girls. Diagonally weighted least squares (WLSMV) estimation was used to account for the ordinal scale of the items [22]. There are several criteria for assessing model fit, such as the chi-square statistic or fit indices. Since the chi-square statistic depends on sample size, it is likely that in large population-based samples, even very small improvements in model fit might become significant [23]. This is why the chi-square statistic was not used for assessing model fit. Instead, model fit was assessed using the Bentler comparative fit index (CFI; [24]), the Tucker–Lewis index (TLI),

where CFI and TLI > 0.90 signifies acceptable fits and > 0.95 signifies good fits, respectively, and the root mean square error of approximation (RMSEA), where an RMSEA < 0.08 indicates an acceptable model fit and < 0.05 a good model fit [25]. Model fit was considered acceptable if CFI ≥ 0.9 and TLI ≥ 0.9, and RMSEA < 0.08. All fit indices were computed from the scaled chi-square statistic (therefore, termed 'scaled CFI', etc.).

Multi-group confirmatory factor analysis (MGCFA) for assessing measurement invariance across age

MGCFA was used to assess measurement invariance of the SDQ across age. A categorisation of the sample into the following age groups was performed: 3–4 years, 5–6 years, 7–8 years, 9–10 years, 11–12 years, 13–14 years and 15–17 years. This categorisation is a trade-off between sufficient sample sizes per age group and homogeneous groups. With this categorisation, the numbers of subjects in each age group within girls or boys were not below 900 and are thus sufficiently large for MGCFA.[2] At the same time, the pooling of two (three, resp.) adjacent ages to form homogeneous age groups is considered to be acceptable.

First, the proposed five-factor model (termed baseline model), in which factor loadings and thresholds varied freely over age groups, was assessed based on fit indices. Configural invariance was assumed if this model had acceptable model fit. Note that the SDQ contains categorical items that are evaluated on an ordinal scale (answer format: untrue, somewhat true, certainly true), and invariance testing with continuous and categorical items differs. According to Muthén and Muthén, in the presence of categorical item responses, thresholds and loadings should be varied in tandem since the item characteristic curves depend on both parameters [27] (see also [28] for the MGCFA methodology with categorical item responses). Thus, weak invariance testing is appropriate for continuous but not for categorical item responses and was not applied for this reason.

If configural invariance held, strong measurement invariance was assessed by comparing the baseline model to a more constrained model (termed strong invariance model), in which all items' factor loadings and thresholds were held equal across age groups. Strong measurement invariance was not established if the difference in the models' CFI indices (ΔCFI) exceeded 0.01 [29, 30]. Note that this is a tolerant criterion and more strict criteria for declaring measurement non-invariance were proposed. Meade et al., for example, proposed declaring measurement non-invariance if ΔCFI ≥ 0.002, which has greater power to detect non-invariance if it is present but also bears a higher risk to falsely declare measurement non-invariance [31]. In this study, we used the less strict criterion (ΔCFI ≥ 0.01) for assuming measurement non-invariance because we want to minimise the risk of incorrectly declaring the SDQ being measurement non-invariant.

If the difference in CFI indices was larger than 0.01 (i.e., strong measurement invariance cannot be assumed), we assessed whether it suffices to constrain the factor loadings and thresholds not for all but only for a few items, usually referred to as partial strong measurement invariance. Partial strong measurement invariance would indicate that specific items function differently on children of different ages but others do not. To identify items for which measurement invariance cannot be assumed, we tested for each item i whether the strong invariance model has a significantly worse fit than a (partial strong invariance) model in which the factor loadings and thresholds are held constant over all age groups, except for the loadings and thresholds of item i that were allowed to freely vary over the age groups. Items with small p values (or equivalently, large score test statistics) can be considered as items for which the constraint of equal loadings and thresholds should be released. We, therefore, first sorted the items according to their p values or equivalently, according to their score test statistics. Then, starting from the strong invariance model, we repeatedly fit a number of models, each time releasing the constraint for one additional item, until we obtained a model which had not a substantially worse fit than the baseline model (i.e., ΔCFI < 0.01).

In contrast to the derivation of centile curves (described in the following paragraph), we did not use weighting factors for (MG)CFA since we do not report any numbers or percentages from these analyses that are supposed to be representative for the population in Germany.

Centile curves for deriving age-specific norms

The LMSP method of centile estimation was used to model centiles of the SDQ total score in dependence on age [32]. This method assumes that for a given age, there is a transformation of the form

$$\widetilde{\mathrm{SDQ}} = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{\mathrm{SDQ}}{\mu}\right)^{\nu} - 1\right] & \text{if } \nu \neq 0 \\ \frac{1}{\sigma}\log\left(\frac{\mathrm{SDQ}}{\mu}\right) & \text{if } \nu = 0 \end{cases},$$

such that the transformed SDQ total score, $\widetilde{\mathrm{SDQ}}$, follows a standard power exponential distribution with power parameter $\tau > 0$. The SDQ score is then said to have a Box–Cox power exponential distribution with parameters μ, σ, ν and τ relating to the location, scale, skewness and kurtosis,

[2] Kline [26], for example, suggests a minimum of 100 subjects per group.

Table 1 Results of MGCFA for assessing measurement invariance across age groups in a specified population

Population	Model	CFI[a]	ΔCFI[d]	TLI[b]	ΔTLI[d]	RMSEA[c]	ΔRMSEA[d]
Boys	Configural invariance	0.925	–	0.915	–	0.049	–
	Strong invariance[e]	0.908	0.017	0.907	0.008	0.051	−0.0020
	Partial strong invariance[e] (items *worries*, *distractible*)	0.916	0.009	0.914	0.001	0.050	−0.0002
Girls	Configural invariance	0.918	–	0.907	–	0.047	–
	Strong invariance[e]	0.898	0.020	0.897	0.009	0.049	−0.0020
	Partial strong invariance[e] (items *somatic*, *worries*, *restless*)	0.910	0.008	0.908	−0.001	0.046	0.0003

[a]Comparable fit index

[b]Tucker–Lewis index

[c]Root mean square error of approximation

[d]Difference in CFI, TLI and RMSEA, respectively, of the baseline model (all parameters estimated freely) and the model with constraints (equal factor loadings and thresholds for all age groups)

[e](Partial) strong measurement invariance is assumed if ΔCFI < 0.01

respectively [32]. Each of the four parameters μ, σ, ν and τ were modelled as smooth non-parametric functions of the exact age. The scores of the subscales take values between 0 and 10, and were assumed to follow a zero-adjusted gamma distribution. Worm plots [33] and Q statistics testing normality of residuals within age groups [34] were used as diagnostic tools to identify possible inadequacies of the fit. In addition, the smoothed centiles were also compared with their empirical counterparts. A weighting factor was used for modelling centiles to correct for deviations in the sample from the population structure in Germany (as on 31 December 2010) with respect to age and region (East/West/Berlin).

Statistical software

All analyses were conducted with the statistical software R, version 3.3.0. CFA and MGCFA models were fit using the function cfa in the R package lavaan (version 0.5-22; [35]). Partial measurement invariance was assessed based on the results of the function lavTestScore of the same package. For modelling the centiles of the SDQ scores, the R package gamlss (version 5.0-2) and relevant functions therein were used [36].

Results

Factor structure and measurement invariance across age groups

SDQ data from 13,423 completed questionnaires (i.e., no missing items; 6810 boys and 6613 girls) were used for assessing the factor structure and measurement invariance. The fit of the five-factor models in the overall study population and within boys and girls was not optimal, and yielded CFI and TLI values below 0.9 (results not shown). The modification indices for all three models suggested that there is strong residual correlation between the items *restless* and *fidgety* of the hyperactivity/inattention subscale. After accounting for residual covariance between these items, the fit considerably improved and yielded acceptable values with CFI = 0.912, TLI = 0.900, RMSEA = 0.051 for the overall study population, CFI = 0.917, TLI = 0.905, RMSEA = 0.053 for boys and CFI = 0.912, TLI = 0.900, RMSEA = 0.050 for girls. The path diagrams for the three models including the factor loadings, thresholds and (co)variances are shown in Online Resource 1.

The models for boys and girls were subsequently specified in the framework of MGCFA to test for configural invariance across all age groups within boys and girls. The configural invariance models yielded an acceptable fit; CFI = 0.925, TLI = 0.915 and RMSEA = 0.049 for boys and CFI = 0.918, TLI = 0.907 and RMSEA = 0.047 for girls (Table 1). This suggests that the proposed five-factor structure of the German SDQ-P is appropriate for the complete age range, 3–17 years.

The strong measurement invariance models (i.e., models with equal item loadings and thresholds across age groups) yielded a substantially worse fit for both boys and girls. The differences in CFI (ΔCFI) exceeded the threshold 0.01. This suggests that the SDQ is not measurement invariant across age groups.

Partial strong measurement invariance was assessed by testing the strong invariance model against a model in which the factor loadings and thresholds of an item may vary freely across age groups. An overview of the score test statistics

Table 2 Items sorted by the score test statistic (in descending order) for assessing partial strong measurement invariance

Population	Item	Subscale	Total score test statistic[a]	*p* value
Boys	Worries[b]	Emotional symptoms	293.7601	< 0.001
	Distractible[b]	Hyperactivity/inattention	282.9750	< 0.001
	Fidgety	Hyperactivity/inattention	253.6173	< 0.001
	Restless	Hyperactivity/inattention	233.6245	< 0.001
	Obeys	Conduct problems	165.1607	< 0.001
	Attends	Hyperactivity/inattention	133.8711	< 0.001
	Somatic	Emotional symptoms	133.6529	< 0.001
	Afraid	Emotional symptoms	121.8226	< 0.001
	Oldbest	Peer problems	109.6697	< 0.001
	Fights	Conduct problems	105.1847	< 0.001
	Reflective	Hyperactivity/inattention	102.0722	< 0.001
	Helpsout	Prosocial behaviour	86.1732	< 0.001
	Bullied	Peer problems	81.3483	< 0.001
	Tantrum	Conduct problems	79.9153	< 0.001
	Clingy	Emotional symptoms	73.0902	< 0.001
	Steals	Conduct problems	68.8376	< 0.001
	Lies	Conduct problems	57.6251	< 0.001
	Shares	Prosocial behaviour	48.9003	< 0.001
	Kind	Prosocial behaviour	33.9919	0.013
	Caring	Prosocial behaviour	29.5748	0.042
	Unhappy	Emotional symptoms	28.8386	0.050
	Popular	Peer problems	27.4548	0.071
	Considerate	Prosocial behaviour	25.0722	0.015
	Loner	Peer problems	24.5483	0.017
	Friend	Peer problems	8.3839	0.972
Girls	Somatic[b]	Emotional symptoms	307.3938	< 0.001
	Worries[b]	Emotional symptoms	225.8184	< 0.001
	Restless[b]	Hyperactivity/inattention	195.3168	< 0.001
	Afraid	Emotional symptoms	190.7193	< 0.001
	Fidgety	Hyperactivity/inattention	164.4548	< 0.001
	Distractible	Hyperactivity/inattention	161.2687	< 0.001
	Obeys	Conduct problems	139.1569	< 0.001
	Clingy	Emotional symptoms	126.8454	< 0.001
	Shares	Prosocial behaviour	102.4874	< 0.001
	Helpsout	Prosocial behaviour	90.9688	< 0.001
	Bullied	Peer problems	85.4670	< 0.001
	Steals	Conduct problems	80.9283	< 0.001
	Oldbest	Peer problems	77.6675	< 0.001
	Reflective	Hyperactivity/inattention	77.6178	< 0.001
	Attends	Hyperactivity/inattention	54.1608	< 0.001
	Lies	Conduct problems	53.0535	< 0.001
	Fights	Conduct problems	50.2452	< 0.001
	Considerate	Prosocial behaviour	44.6674	< 0.001
	Tantrum	Conduct problems	41.7017	< 0.001
	Friend	Peer problems	33.1839	0.016
	Popular	Peer problems	29.1721	0.046
	Loner	Peer problems	26.8320	0.008
	Unhappy	Emotional symptoms	24.3841	0.143
	Kind	Prosocial behaviour	22.0147	0.231
	Caring	Prosocial behaviour	21.3789	0.261

[a]The score test was used to compare the strong invariance model to a partial strong invariance model with the factor loadings and thresholds varying freely over the age groups for the specified item and equal factor loadings and thresholds for all other items [6 (no. freely varying factor loadings) + 12 (no. freely varying thresholds) = 18 degrees of freedom; *df*]. For the first item of each subscale, loadings are fixed to 1 to achieve identifiability, thus *df* = 12 for these items. Large test statistics suggest that the item's factor loadings and/or thresholds differ between age groups and that the item is thus not measurement invariant across age

[b]Items for which factor loadings and thresholds were freed in the final chosen partial strong invariance model

and *p* values of all 25 items is given in Table 2. For boys, the strongest evidence against measurement invariance (i.e., the largest test statistic, or equivalently the smallest *p* value) was obtained for the item *worries* of the emotional symptoms subscale and items *distractible*, *fidgety* and *restless* of the hyperactivity/inattention subscale. For the item *distractible*, there was a non-linear change in the thresholds (see figure in Online Resource 2): The thresholds decreased within childhood and then increased again showing that 3–6 and 15–17-year-olds are less likely to become distracted than 7–14-year-olds. The item *worries* is less often endorsed by parents of younger children (3–10 years) since thresholds decrease constantly during these ages (Online Resource 2). The thresholds for the item *fidgety* were larger for older boys (results not shown). This shows that parents of younger children endorse this item more than parents of older children or adolescents. Releasing the equality constraints for the items *worries* and *distractible* yielded an acceptable partial strong invariance model with CFI = 0.916, TLI = 0.914 and RMSEA = 0.050. This model was not substantially worse than the baseline model (ΔCFI = 0.009 < 0.01; see Table 1 and Online Resource 2 for details). Partial strong measurement invariance can thus be established for boys.

For girls, the biggest problems were observed also for items of the emotional symptoms subscale (*somatic*, *worries*, *afraid*) and the hyperactivity/inattention subscale (*restless*, *fidgety*), as these items yield the largest score test statistics (see Table 2). An inspection of the age-group-specific thresholds for the items *worries* and *somatic* shows that parents of older children and adolescents are more likely to report that their child worries or has headaches, stomach-aches and sickness, respectively, than parents of younger children (see figure in Online Resource 2). Parents of younger children, in contrast, are more likely to report that their child has many fears or is easily scared (item *afraid*; results not shown), is constantly fidgeting or squirming (item *fidgety*; results not shown) and is restless or overactive (item *restless*; Online Resource 2). Relaxing the equality constraints for the three most problematic items, *somatic*, *worries* and *restless*, yielded a partial strong invariance model that was not substantially worse than the baseline model (CFI = 0.910, TLI = 0.908, RMSEA = 0.046; ΔCFI = 0.008 < 0.01; cf. Table 1). Partial strong invariance can, therefore, be established also for girls.

To conclude, the five-factor structure of the SDQ can be validated in all age groups. The SDQ and the subscales are thus applicable also in children younger than 6 years that have not been investigated in studies on the German SDQ-P so far. The results obtained from MGCFA for both boys and girls suggest that strong measurement invariance cannot be assumed for the emotional symptoms subscale and the hyperactivity/inattention subscale, while for the other three subscales, measurement invariance might be assumed. This supports the use of age-specific norms at least for the emotional symptoms subscale and the hyperactivity/inattention subscale. Age-specific norms should also be used for the total difficulties score since this is the sum of the subscale scores.

Age-specific norm values for the SDQ total difficulties score

The data of completed questionnaires (n = 13,423; 90.5%) and incomplete questionnaires with no more than two missing items per subscale (n = 1054; 7.1%) were used to derive age-specific reference values for the German SDQ-P. This makes up 97.6% of all questionnaires. Only 2.4% (n = 358 in total; 165 girls; 193 boys) of the questionnaires had to be excluded from the computation of reference values due to too many incomplete items.

The age-specific 5th, 10th, 20th, 50th, 80th, 90th, 95th centile curves for the SDQ total difficulties score are shown in Fig. 1 separately for boys and girls. The concrete values for the 50th, 80th, 90th and 95th centiles of the SDQ total difficulties score are specified in Table 3 for children of age 3–17.

Figure 1 supports a dependency of the centiles on age. In particular, the upper centiles that are commonly used for defining abnormal behaviour show a strong dependency on age, and the dependency is stronger for higher centiles. The upper (i.e., 80th, 90th, 95th) centiles have their maxima at about 10 years for both boys and girls. This implies that a total difficulties score of 17, for example, is more indicative for abnormal behaviour for a 17-year-old boy than it is for a 10-year-old boy because about 10% of the 10-year-old boys score equal or higher than 12, while only about 5% of the 17-year-old boys have such a high or an even higher score. Or using the notation of Goodman et al., for the 10-year-old the SDQ score of 17 is considered high (90th–95th centile), while for a 17-year-old the same score is considered very high (> 95th centile) [7].

The median and the lower centiles do not show any differences during childhood. For boys, the median score is constantly high at 8.5 up to the age of 13 years, gets smaller with increasing age and has its minimum for 17-year-olds with about 6–7 points. For girls, the median slightly decreases from approx. 8 to 7 at the beginning of a girl's lifetime, from 6 to 13 years it is constant at 7 and then slightly decreases at about 1 score. The lower (i.e., 5th, 10th, 20th) centiles show a similar dependency on age.

Age-specific norm values for subscales

We computed the 80th, 90th and 95th centiles of the subscales since these centiles are frequently used for screening children and adolescents with hyperactivity/inattention

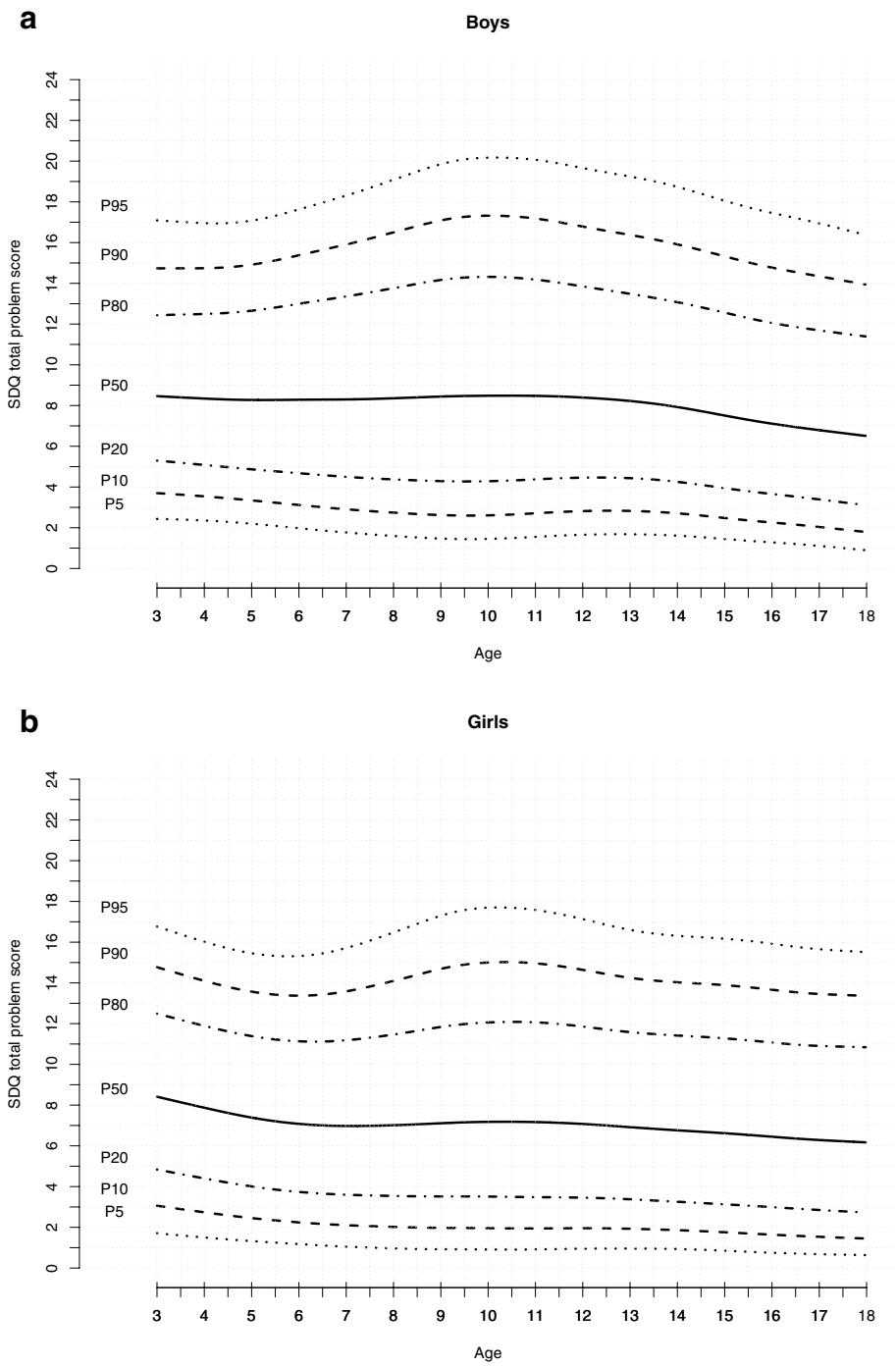

Fig. 1 Age-specific norm values for the total difficulties score of the German SDQ-P. The plots show the 5th (P5), 10th (P10), 20th (P20), 50th (P50), 80th (P80), 90th (P90), 95th (P95) centiles of the SDQ total difficulties score in dependence on age in the German study population including 3–17-year-old boys (**a**) and girls (**b**)

problems, conduct problems, emotional problems, peer problems or deficits in prosocial behaviour, respectively. In addition to those, the median score (50th centile) was computed in order to assess changes of the central tendency with age. The age-specific centiles for the five subscale scores are shown in Fig. 2 separately for boys (a, c, e, g, i) and girls (b, d, f, h, j).

The results for the emotional symptoms subscale and the hyperactivity/inattention subscale suggest the use of age-specific norms for these two subscales as differences in the centiles over age can be observed for the two subscales (Fig. 2a–d). Concrete values of the age-specific 50th, 80th, 90th, 95th centiles are given in Table 4. Note that differences in the 80th, 90th and 95th centiles are remarkable, while the median does not vary that much with age. There is a tendency of higher scores around the age of 10 years, in particular for the emotional symptoms subscale. In girls, higher scores are again reached in late adolescence,

Table 3 Age-specific norm values for the SDQ total difficulties score of the German SDQ-P for boys and girls

Age (in years)	Boys				Girls			
	P50	P80	P90	P95	P50	P80	P90	P95
3[a]	8.46	12.44	14.73	17.09	8.41	12.49	14.77	16.76
4[a]	8.34	12.50	14.74	16.96	7.87	11.89	14.10	16.02
5	8.27	12.66	14.92	17.08	7.39	11.39	13.57	15.46
6	8.28	12.99	15.38	17.63	7.08	11.13	13.37	15.32
7	8.30	13.36	15.91	18.31	6.98	11.19	13.59	15.71
8	8.36	13.76	16.50	19.09	7.01	11.47	14.10	16.48
9	8.44	14.16	17.09	19.85	7.11	11.84	14.69	17.30
10	8.48	14.32	17.32	20.17	7.18	12.06	15.00	17.69
11	8.48	14.18	17.18	20.06	7.17	12.06	14.96	17.58
12	8.39	13.85	16.79	19.66	7.07	11.86	14.64	17.13
13	8.22	13.48	16.38	19.24	6.92	11.58	14.25	16.60
14	7.93	13.07	15.91	18.74	6.76	11.42	14.03	16.31
15	7.51	12.56	15.32	18.05	6.62	11.28	13.89	16.17
16	7.11	12.06	14.77	17.46	6.45	11.07	13.67	15.93
17	6.79	11.69	14.34	16.94	6.29	10.91	13.46	15.65

Values were taken from smoothed centile curves (cf. Fig. 1) based on the LMSP method. SDQ scores are regarded being 'close to average' if they are equal or smaller than the 80th centile (P80), 'slightly raised' if they lie between the 80th and 90th centiles (P80–P90), 'high' if they lie between the 90th and 95th centiles (P90–P95) and 'very high' if they exceed the 95th centile (P95)

[a]Not based on the early-year version of the SDQ-P

while for boys the centiles are steadily smaller with age. The centiles for the hyperactivity/inattention subscale are smaller for ages above 10 years. In contrast to the peak at about 10 years observed for the emotional symptoms subscale, the centiles for the hyperactivity/inattention subscale remain nearly the same for 3–10-year-olds, for girls the highest scores are even obtained for 3-year-olds.

In contrast to the emotional symptom and the hyperactivity/inattention subscales, the differences in the subscales on conduct problems, peer problems and prosocial behaviour are rather small (Fig. 2e–j, Online Resource 3). In particular, there are hardly any differences for the conduct problem subscale, suggesting that scores of the conduct problem subscale are comparable across different age groups. Very small differences across age can be observed in the upper centiles for the peer problem subscale and the prosocial behaviour subscale. Note that the 'jump' of the median to exactly zero in Fig. 2h is an artefact which arises from the fact that a mixed discrete–continuous distribution (zero-adjusted gamma distribution) was used. Changes in the centiles of the peer problem subscale were similar for boys and girls but slightly more remarkable for girls than for boys. The upper centiles for girls take slightly smaller values from 3 up to approx. 6 years, increase from 6 to 11 years and decrease again. The same can be observed for boys, but the changes in the centiles are smaller. Note that we inverted the scores of the prosocial behaviour subscale: higher scores obtained on this subscale indicate worse prosocial behaviour. For the prosocial behaviour subscale, the centiles first slightly decrease from 3 to approx. 9 years and then increase again.

Table 5 shows the norm values not specific to age for the subscales on conduct problems, peer problems and prosocial behaviour, respectively. Note that the subscales have a limited number of values (cf. Online Resource 3). Thus, it is common that there are no values that exactly correspond to the 80th, 90th, and 95th centiles, as was already noted in other papers (e.g., [12, 13]). In some cases, there is even a large deviation between the exact centiles and the 80th, 90th, and 95th centiles, as seen from Table 5.

Discussion

With the present study, we validated the five-factor structure of the German SDQ-P across the entire age range of 3–17 years. Our results are in accordance with prior studies on the validation of the factor structure for the German SDQ-P, namely that the German SDQ-P can be used for children and adolescents [11–13]. Since we were the first to also include children younger than 6 years from a nationally representative sample, we additionally conclude that the SDQ and the subscales can be derived from parent ratings of preschoolers, as the study by Klein et al. on a small regional sample has suggested [10].

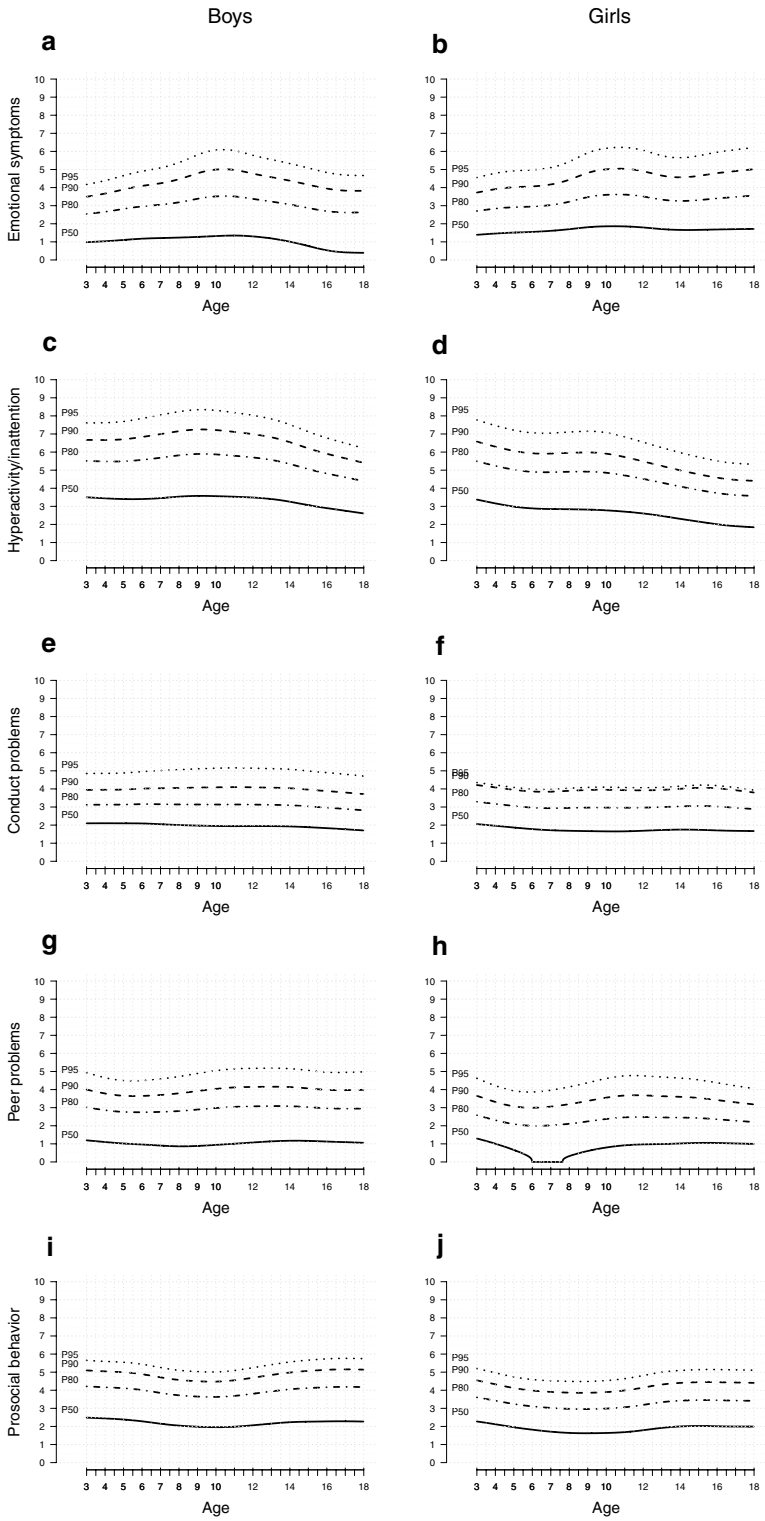

Fig. 2 Age-specific norm values for the subscale scores of the German SDQ-P. The plots show the 50th, 80th, 90th, 95th centiles of the subscale score in dependence on age in the German study population including 3–17-year-old boys (**a**, **c**, **e**, **g**, **i**) and girls (**b**, **d**, **f**, **h**, **j**). The prosocial behaviour subscale was inverted such that lower values indicate better prosocial behaviour

SDQ scores are often interpreted irrespectively of the child's age. The underlying assumption is that identical scores represent the same level of the construct measured by the SDQ for children of different ages. Both MGCFA and centile curves revealed that the SDQ is not measurement invariant across age, although the factor structure could be confirmed across the entire age range of the present study. The results of this study help with identifying children with abnormal behaviour and rating the severity of abnormality, while explicitly taking the developmental course of SDQ problems into account. For example, we showed that no more than 5% of the 4-year-old boys in Germany are

Table 4 Age-specific norm values for the emotional symptoms subscale and the hyperactivity/inattention subscale of the German SDQ-P for boys and girls

Age (in years)	Emotional symptoms								Hyperactivity/inattention							
	Boys				Girls				Boys				Girls			
	P50	P80	P90	P95	P50	P80	P90	P95	P50	P80	P90	P95	P50	P80	P90	P95
3[a]	0.98	2.54	3.49	4.17	1.39	2.71	3.72	4.55	3.50	5.51	6.67	7.61	3.37	5.49	6.58	7.77
4[a]	1.03	2.66	3.67	4.39	1.46	2.84	3.91	4.79	3.44	5.48	6.67	7.62	3.16	5.24	6.30	7.46
5	1.11	2.82	3.90	4.66	1.51	2.91	4.02	4.92	3.40	5.49	6.71	7.69	2.99	5.03	6.07	7.21
6	1.17	2.96	4.09	4.90	1.55	2.95	4.07	4.98	3.41	5.57	6.83	7.86	2.89	4.91	5.93	7.06
7	1.21	3.05	4.24	5.09	1.61	3.04	4.17	5.11	3.46	5.69	6.99	8.05	2.85	4.89	5.92	7.05
8	1.24	3.19	4.47	5.39	1.70	3.22	4.44	5.44	3.54	5.82	7.15	8.23	2.84	4.91	5.96	7.11
9	1.27	3.38	4.79	5.80	1.81	3.47	4.81	5.92	3.58	5.89	7.25	8.34	2.82	4.92	5.98	7.15
10	1.32	3.51	5.00	6.07	1.86	3.60	5.01	6.17	3.56	5.87	7.22	8.30	2.78	4.86	5.91	7.07
11	1.35	3.51	4.98	6.04	1.85	3.61	5.04	6.22	3.53	5.80	7.11	8.18	2.71	4.71	5.73	6.84
12	1.30	3.38	4.77	5.79	1.79	3.51	4.90	6.06	3.49	5.70	6.99	8.03	2.61	4.52	5.48	6.54
13	1.19	3.23	4.57	5.55	1.71	3.34	4.67	5.77	3.40	5.57	6.82	7.83	2.47	4.31	5.23	6.25
14	1.01	3.06	4.38	5.33	1.66	3.26	4.57	5.65	3.26	5.34	6.54	7.52	2.31	4.10	4.99	5.97
15	0.77	2.87	4.16	5.08	1.66	3.31	4.65	5.76	3.07	5.06	6.20	7.13	2.16	3.90	4.77	5.72
16	0.54	2.70	3.95	4.84	1.69	3.40	4.80	5.96	2.90	4.82	5.91	6.78	2.01	3.74	4.58	5.51
17	0.42	2.62	3.83	4.69	1.70	3.48	4.91	6.09	2.76	4.61	5.65	6.49	1.91	3.63	4.46	5.38

Values taken from smoothed centile curves (cf. Fig. 2a–d). SDQ subscale scores might be regarded being 'close to average' if they are equal or smaller than the 80th centile (P80), 'slightly raised' if they lie between the 80th and 90th centiles (P80–P90), 'high' if they lie between the 90th and 95th centiles (P90–P95) and 'very high' if they exceed the 95th centile (P95)

[a]Not based on the early-year version of the SDQ-P

Table 5 Norm values (not age-specific) for the conduct problems subscale, the peer problems subscale and the prosocial behaviour subscale of the German SDQ-P for boys and girls

Subscale	Boys				Girls			
	P50	P80	P90	P95	P50	P80	P90	P95
Conduct problems	2.00 (63.3)[b]	3.00 (81.1)[b]	4.00 (91.8)[b]	5.00 (96.4)[b]	1.67 (47.9)[b]	3.00 (87.2)[b]	3.75 (87.4)[b]	4.00 (95.0)[b]
Peer problems	1.00 (56.9)[b]	3.00 (86.3)[b]	4.00 (93.3)[b]	5.00 (96.9)[b]	1.00 (62.4)[b]	2.00 (80.0)[b]	3.75 (90.0)[b]	4.00 (95.4)[b]
Prosocial behaviour[a]	2.00 (53.5)[b]	4.00 (86.1)[b]	4.00 (86.1)[b]	5.00 (95.0)[b]	1.67 (45.0)[b]	3.00 (82.8)[b]	4.00 (92.4)[b]	5.00 (97.4)[b]

SDQ subscale scores might be regarded being 'close to average' if they are equal or smaller than the 80th centile (P80), 'slightly raised' if they lie between the 80th and 90th centiles (P80–P90), 'high' if they lie between the 90th and 95th centiles (P90–P95) and 'very high' if they exceed the 95th centile (P95)

[a]The prosocial behaviour subscale was inverted such that lower values indicate better prosocial behaviour

[b]Exact centiles, i.e., the (weighted) fraction of participants with subscale score not exceeding the reference value

expected to have an SDQ total score exceeding 16.96 (95th centile). A 4-year-old boy with an SDQ total score of 18, say, could accordingly be rated as having a clinically relevant behaviour that is abnormal for boys of that age. Being aware of those differences across age groups is critical. Without knowledge of differences in the SDQ scales between age groups, services might be denied to children of specific ages because their SDQ scores are below a clinical cutoff despite high levels of impairment [28]. Further, as was noted by Bowen and Masa "Researchers might draw erroneous conclusions about relationships among social, emotional, or behavioural constructs and outcomes for subgroups [and] Their conclusions could translate into guidelines for intervention that are inappropriate for some clients" [28]. We, therefore, strongly encourage the use of age-specific norms for the SDQ total difficulties score, as well as for the emotional symptoms subscale and the hyperactivity/inattention subscale. Norms that are not specific to age might be used for the other three subscales since these were shown to be measurement invariant across age.

Note that the KiGGS study population is far larger than that of most of the existing studies which derive norms for

the SDQ or assess psychometric properties of the SDQ [5]. The number of subjects of each age is sufficiently large in KiGGS for investigations on measurement invariance across the complete age range. The present study allows detailed analyses on measurement invariance of the German SDQ-P across all ages, from early childhood to late adolescence, as well as the establishment of age-specific norms. This is in contrast to the existing studies [5]. Neither of these studies covered the complete age range from 3 to 17 years. The focus on a narrow age range prevents an assessment of SDQ measurement invariance across the complete age range. Further, the existing studies did not include a large number of subjects of the same age, such that children of different ages were allocated to an age group. In particular, age groups that cover a wide age range might be too heterogeneous and differences within age groups (e.g., in norm values) are concealed. Rothenberger et al., for example, reported no differences across age from their results of MGCFA in contrast [11]. In their MGCFA, they subsumed children aged 7–10 years in one group and children of age 11–16 years in another group. The centile curves we have derived show that there are considerable differences within each of the two age groups. Further, our studies show that centiles do not increase or decrease linearly with age but that there is a non-linear change with a "peak" at the age of around 10 years indicating that the variability of SDQ scores is largest for children at the age of around 10 years. Younger or older children have narrower normal ranges in contrast. These findings suggest that the categorisation used by Rothenberger et al. is too rough to detect any differences across age since aggregated values of 7–10-year-olds and 11–16-year-olds are similar. Note that the sample from the BELLA study which, as used in their studies, is a subsample of our study population, supporting our theory that the different conclusions are not based on sample differences but are related to the subsuming of heterogeneous groups to one broad age group.

In principle, we might have also used centiles that are observed in each age group as norm values. However, due to a limited number of subjects within each year, there is a large variation in centiles, in particular for extreme centiles like the 95th or the 90th centiles which are of special interest in the context of identifying children with abnormal behaviour. Another disadvantage of this approach is that the exact age of survey participants would be neglected when computing centiles for each year. The reference value for a boy who had just had his 12th birthday would ideally be derived from children of exactly the same age, rather than from 12-year-old boys who are just turning 13. Smoothed centile curves account for random variations in the centiles and they reflect changes in the course of life. Centile curves have frequently been used in population-based studies to flexibly model the centiles of specific variables in dependence on age. They have become an established tool for measurements related to growth and development in the context of paediatrics. To our knowledge, this method has rarely been used in psychology, and in particular not in the context of age-specific norm values for the SDQ. Note that the presented centiles which might be used as cutoff values for identifying children with clinically relevant behaviour do not require the subscale score or SDQ total score to be an integer. In the presence of incomplete questionnaires, we recommend comparing real-valued scores to the centiles since real-valued scores are more precise.

When using age-specific norms, one should, however, be aware that the categorisation of SDQ scores as abnormal (90th percentile) or borderline (80th percentile) is based on the prevalence of mental health problems of 10–20% for children and adolescents, and stems from the entire age range rather than from specific age subgroups. Goodman advises using cutoffs based on knowledge of the prevalence in the general population [3]. Further, he notes that it "may be appropriate to adjust cutoffs for age and gender" ([3], p. 585). Age-specific prevalence rates based on nationally representative samples have, however, not been reported so far, and our nationally representative sample does not allow estimating the prevalence of mental health problems in a reliable way. Future studies are needed to address this issue. In their latest manuscript on this issue, Goodman et al. proposed using the 80th, 90th and 95th centiles as cutoffs for defining the scores as 'close to average' (< 80th centile), 'slightly raised' (80th–90th centile), 'high' (90th–95th centile) and 'very high' (> 95th centile) instead of categorising individuals as abnormal [7]. This categorisation does not depend on the prevalence of psychopathological abnormalities. It is thus applicable in populations with a different prevalence of psychopathological abnormalities, such as boys and girls or children of different ages. We, therefore, recommend using this interpretation of SDQ scores as long as there is no knowledge on the age-specific prevalence of mental health problems.

Conclusion

We used data from a large nationally representative survey (KiGGS) for providing norm values and validating the German parent version of the SDQ (SDQ-P) across the complete age range. For the first time, evidence was provided that the German SDQ-P is a valid screening instrument also for preschoolers. Moreover, we showed that for neither boys nor girls the SDQ-P is measurement invariant across age. Results from both MGCFA and centile curves showed that the absence of measurement invariance is attributable to a different answer behaviour to some items of the emotional symptoms subscale (item "worries" for both boys and girls

and item "somatic" for girls) and the hyperactivity/inattention subscale (item "distractible" for boys and "restless" for girls). For screening mental health problems, we, therefore, propose using age-specific norms and cutoff values for the SDQ total difficulties score and for the subscales on hyperactivity/inattention and emotional symptoms. In contrast to that, we propose using generic norms and cutoff values for the subscales on conduct problems, peer problems and prosocial behaviour since these subscales were shown to be measurement invariant across age. Norm values for the SDQ and its subscales were derived from data of a large nationally representative sample and are provided along with this paper.

Acknowledgements The authors thank Angelika Schaffrath Rosario for helpful comments on centile curves and Carol Wallis for language corrections.

Funding Funding was provided by Bundesministerium für Gesundheit.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Cicchetti D (2006) Development and psychopathology. In: Cicchetti D, Cohen DJ (eds) Developmental psychopathology, Vol 1: theory and method, 2nd edn. Wiley, Hoboken
2. Hudziak JJ et al (2007) A dimensional approach to developmental psychopathology. Int J Methods Psychiatr Res 16(Suppl 1):S16–S23
3. Goodman R (1997) The Strengths and Difficulties Questionnaire: a research note. J Child Psychol Psychiatry 38(5):581–586
4. Goodman R et al (2000) Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br J Psychiatry 177:534–539
5. Stone LL et al (2010) Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: a review. Clin Child Fam Psychol Rev 13(3):254–274
6. Barkmann C, Schulte-Markwort M (2004) Prävalenz psychischer Auffälligkeit bei Kindern und Jugendlichen in Deutschland - ein systematischer Literaturüberblick. Psychiatr Prax 31(6):278–287
7. Goodman A, Lamping DL, Ploubidis GB (2010) When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): data from British parents, teachers and children. J Abnorm Child Psychol 38(8):1179–1191
8. Becker A et al (2006) Psychopathological screening of children with ADHD: Strengths and Difficulties Questionnaire in a pan-European study. Eur Child Adolesc Psychiatry 15(Suppl 1):56–62
9. Becker A et al (2004) Validation of the parent and teacher SDQ in a clinical sample. Eur Child Adolesc Psychiatry 13(Suppl 2):11–16
10. Klein AM et al (2013) Psychometric properties of the parent-rated SDQ in preschoolers. Eur J Psychol Assess 29(2):96–104
11. Rothenberger A et al (2008) Psychometric properties of the parent strengths and difficulties questionnaire in the general population of German children and adolescents: results of the BELLA study. Eur Child Adolesc Psychiatry 17(Suppl 1):99–105
12. Woerner W et al (2002) Normierung und Evaluation der deutschen Elternversion des Strengths and Difficulties Questionnaire (SDQ): Ergebnisse einer repräsentativen Felderhebung. Zeitschrift fur Kinder- und Jugendpsychiatrie und Psychotherapie 30(2):105–112
13. Woerner W, Becker A, Rothenberger A (2004) Normative data and scale properties of the German parent SDQ. Eur Child Adolesc Psychiatry 13(Suppl 2):3–10
14. Stone LL et al (2010) Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: a review. Clin Child Fam Psychol Rev 13(3):254–274
15. Goodman A, Goodman R (2011) Population mean scores predict child mental disorder rates: validating SDQ prevalence estimators in Britain. J Child Psychol Psychiatry 52(1):100–108
16. Hölling H et al (2007) Behavioral problems in children and adolescents. First results of the German Health Interview and Examination Survey for Children and Adolescents (KiGGS) [Verhaltensauffälligkeiten bei Kindern und Jugendlichen. Erste Ergebnisse aus dem Kinder- und Jugendgesundheitssurvey (KiGGS)]. Bundesgesundheitsblatt 50(5–6):784–793
17. Kurth BM (2007) The German Health Interview and Examination Survey for Children and Adolescents (KiGGS): an overview of its planning, implementation and results taking into account aspects of quality management. Bundesgesundheitsblatt 50(5–6):533–546
18. Rothenberger A et al (2008) Psychometric properties of the parent Strengths and Difficulties Questionnaire in the general population of German children and adolescents: results of the BELLA study. Eur Child Adolesc Psychiatry 17(Suppl 1):99–105
19. Hölling H et al (2012) Die KiGGS-Studie. Bundesweit repräsentative Längs- und Querschnittstudie zur Gesundheit von Kindern und Jugendlichen im Rahmen des Gesundheitsmonitorings am Robert Koch-Institut. Bundesgesundheitsblatt 55(6):836–842
20. Kamtsiuris P, Lange M, Rosario AS (2007) Der Kinder- und Jugendgesundheitssurvey (KiGGS): Stichprobendesign, Response und Nonresponse-Analyse. Bundesgesundheitsblatt 50(5):547–556
21. Kurth B-M et al (2008) The challenge of comprehensively mapping children's health in a nation-wide health survey: design of the German KiGGS-Study. BMC Public Health 8:196
22. Muthén B (1984) A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 49(1):115–132
23. Schermelleh-Engel K, Moosbrugger H, Müller H (2003) Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods Psychol Res Online 8(2):23–74
24. Bentler PM (1990) Comparative fit indexes in structural models. Psychol Bull 107(2):238–246
25. Schreiber JB et al (2006) Reporting structural equation modeling and confirmatory factor analysis results: a review. J Educ Res 99(6):323–338
26. Kline T (2005) Psychological testing: a practical approach to design and evaluation. Sage Publications, Inc, Thousand Oaks, CA
27. Muthén L, Muthén B (1998–2010) Mplus user's guide, 6th edn. Muthén & Muthén, Los Angeles

28. Bowen N, Masa R (2015) Conducting measurement invariance tests with ordinal data: a guide for social work researchers. J Soc Soc Work Res 6(2):229–249
29. Chen FF (2007) Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Model 14(3):464–504
30. Cheung GW, Rensvold RB (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Model 9(2):233–255
31. Meade AW, Johnson EC, Braddy PW (2008) Power and sensitivity of alternative fit indices in tests of measurement invariance. J Appl Psychol 93(3):568–592
32. Rigby RA, Stasinopoulos DM (2004) Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution. Stat Med 23(19):3053–3076
33. Buuren SV, Fredriks M (2001) Worm plot: a simple diagnostic device for modelling growth reference curves. Stat Med 20(8):1259–1277
34. Royston P, Wright E (2000) Goodness-of-fit statistics for age-specific reference intervals. Stat Med 19(21):2943–2962
35. Rosseel Y (2012) lavaan: an R package for structural equation modeling. J Stat Softw 48(2):1–36
36. Stasinopoulos DM, Rigby RA (2007) Generalized additive models for location scale and shape (GAMLSS) in R. J Stat Softw 23(7):1–46