A Confirmatory Approach to Examining the Factor Structure of the Strengths and Difficulties Questionnaire (SDQ)

A Large Scale Cohort Study

Niclasen, Janni; Skovgaard, Anne Mette; Andersen, Anne-Marie Nybo; Sømhovd, Mikael Julius; Obel, Carsten

Published in:
Journal of Abnormal Child Psychology

DOI:
10.1007/s10802-012-9683-y

Publication date:
2013

Document version
Early version, also known as pre-print

J Abnorm Child Psychol
DOI 10.1007/s10802-012-9683-y

A Confirmatory Approach to Examining the Factor Structure of the Strengths and Difficulties Questionnaire (SDQ): A Large Scale Cohort Study

Janni Niclasen · Anne Mette Skovgaard · Anne-Marie Nybo Andersen · Mikael Julius Sømhovd · Carsten Obel

Abstract The aim of this study was to examine the factor structure of the Strengths and Difficulties Questionnaire (SDQ) using a Structural Confirmatory Factor Analytic approach. The Danish translation of the SDQ was distributed to 71,840 parents and teachers of 5–7 and 10–12-year-old boys and girls from four large scale cohorts. Three theoretical models were examined: 1. a model with five first order factors (i.e., hyperactivity/inattention, conduct, emotional, peer problems and prosocial), 2. a model adding two internalising and externalising second order factors to model 1, and 3. a model adding a total difficulties second order factor to model 1. Model fits were evaluated, multi-group analyses were carried out and average variance extracted (AVE) and composite reliability (CR) estimates were examined. In this general population sample, low risk sample models 1 and 2 showed similar good overall fits. Best model fits were found when two positively worded items were allowed to cross load with the prosocial scale, and cross loadings were allowed for among three sets of indicators. The analyses also revealed that model fits were slightly better for teachers than for parents and better for older children than for younger children. No convincing differences were found between boys and girls. Factor loadings were acceptable for all groups, especially for older children rated by teachers. Some emotional, peer, conduct and prosocial subscale problems were revealed for younger children rated by parents. The analyses revealed more internal consistency for older children rated by teachers than for younger children rated by parents. It is recommended that model 1 comprising five first order factors, or alternatively model 2 with additionally two internalising/externalising second order factors, should be used when employing the SDQ in low risk epidemiological samples.

Keywords Strengths and difficulties questionnaire · SDQ · Psychometric properties · Factor structure · Confirmatory factor analysis · CFA · CR reliability · AVE reliability · Psychopathology · Mental health · Children · Adolescents · Cohort · Questionnaire

J. Niclasen (✉) · M. J. Sømhovd
Department of Psychology, University of Copenhagen,
Øster Farimagsgade 2A,
1353 Copenhagen K, Denmark
e-mail: janni.niclasen@psy.ku.dk

A. M. Skovgaard
Child and Adolescent Psychiatric Centre Glostrup,
Copenhagen University Hospital,
Copenhagen, Denmark

A.-M. N. Andersen
Department of Public Health, University of Copenhagen,
Copenhagen, Denmark

C. Obel
Department of Public Health, University of Aarhus,
Bartholins Allé 2, building 1260, room 126,
8000 Aarhus C, Denmark

Published online: 25 September 2012

The Strengths and Difficulties Questionnaire (SDQ) was developed by Goodman in the mid-1990's as a screening instrument aimed to cover the most prevalent areas of psychopathology in children and adolescents and designed to correspond to the diagnostic categories recognised by the two major diagnostic classification systems, i.e., the International Classification of Diagnosis (ICD-10) (World Health Organisation 1993) and the Diagnostic and Statistical Manual (DSM-IV) (American Psychiatric Association 1994) (Goodman 1994). The 25 SDQ items ask about five distinct domains of psychological adjustment among children and adolescents namely: hyperactivity/inattention, emotional symptoms, conduct problems, peer problems and prosocial behaviours. Apart from the five prosocial items, five problem items are also positively worded in order to enhance acceptability of the questionnaire in the

general population where the majority of children experience relatively few psychopathological difficulties (Goodman 1997; Goodman and Scott 1999).

The factor structure of the 25 SDQ items has been extensively assessed in different cultural settings by means of exploratory factor analysis (EFA) and most studies have been able to confirm the five factor structure (Goodman 2001; Koskelainen et al. 2000; Niclasen et al. 2012). However, as the development of the SDQ was theory driven and since it is assumed that the 25 items reflect five underlying latent dimensions, it seems more appropriate to validate the five scales by means of confirmatory factor analysis (CFA). CFA constitutes the measurement part of structural equation modeling (SEM). It is a technique that analyses measurement models in which both the number of factors and their corresponding indicators are explicitly specified a priori. Relatively few studies have employed structural confirmatory methods in relation to the SDQ and their results vary (Sanne et al. 2009; Van et al. 2008). Thus, some studies have found support for a five-factor model (Palmieri and Smith 2007; Sanne et al. 2009; Van et al. 2008) and others for a three-factor solution (Dickey and Blumberg 2004; Goodman et al. 2010). A study by Goodman et al. (2010) found a three-factor model (internalising/externalising/prosocial) to have a better fit in a low risk epidemiological sample of 5–16-year-olds, but that a five factor model was superior in high risk samples.

While one central issue is concerned with whether SDQ items are truly valid indicators of the proposed five behavioural domains or whether an even simpler structure would be superior, another key issue concerns the impact of the positively worded items. The inclusion of these items was originally intended to increase the acceptability of the SDQ to respondents, making it particularly suitable for use in non-clinical, epidemiological studies. The disadvantage however is, as several studies have pointed out, that positively worded items can confound the factor structure (Goodman 2001; Palmieri and Smith 2007). One study which included proxy data from custodial grandmothers found that a model which contained a positive construct method factor fitted the data better than the three- and five-factor models (Palmieri and Smith 2007). Similarly, a Norwegian study using self-rating data also found a significant improvement of the model fit by introducing a positive construct factor (Van et al. 2008). On the other hand, Sanne et al. (2009) did not find support for a positive construct factor for parent and teacher proxy data.

Thus, the advantages of the structural confirmatory methods are that they provide a comprehensive means for assessing and modifying theoretical models and therefore have a great potential for furthering theory development. The aims of the present paper are three fold. First, to examine how well three overall theoretical models fit data: Model 1. a five factor model (*hyperactivity/inattention, emotional, conduct, peer problems* and *prosocial*); Model 2. a five factor model with 2 s order factors (*internalising*/*externalising*); and Model 3. a five factor model with one latent *total difficulties* factor (Fig. 1). The three theoretical models are included as Goodman found the internalising/externalising model to have better overall fit as compared to the five-factor model in a low risk sample but did not test whether these two models were superior to the original proposed model with a total difficulties second order factor (Goodman et al. 2010). The models are here examined separately for parent ratings and teacher ratings, separately for both 5–7- and 10–12-year-old children and separately for boys and girls. Secondly, after examining the overall model fits, multi-group analyses are carried out in order to test for the presence of multi-group invariance, and to investigate in what ways the groups differ. Thirdly, two measures of reliability, average variance extracted (AVE) and composite reliability (CR), are examined.

Materials and Method

Samples

Data from the four population based, large scale birth cohorts, the Copenhagen Child Cohort (CCC2000), the Danish National Birth Cohort (DNBC), the Danish National Institute of Social Research's (DNISR) and the Aarhus Birth Cohort (ABC) were included in the present study (Table 1). Teacher ratings were available for the ABC and CCC2000 cohorts. The parent samples all had a small overrepresentation of boys whereas the opposite was true for the teacher samples and in all parent samples the questionnaires were mainly filled in by the mothers. As no differences in any analyses were found between the 5- and 7-year samples these were pooled for all analyses presented below and are denoted as younger children. In this way, the parent sample included a total of 63,615 ratings whereas the teacher samples added up to a total of 8,225 ratings.

Loss to follow up varied between the cohorts and various reasons may be responsible for these different response rates (Table 1). One explanation for the relatively low response rate of the DNBC could, for example, be that a large number of general practitioners refused to inform the pregnant women of the study. Similar for all samples, however, was that compared to the background population the samples were underrepresented regarding low socioeconomic resources (education, occupation, income and civil status), parents who were not born in Denmark; younger mothers; parents living separately at the time of birth; and changed family composition in the first 5 years of life (Aarhus Birth Cohort

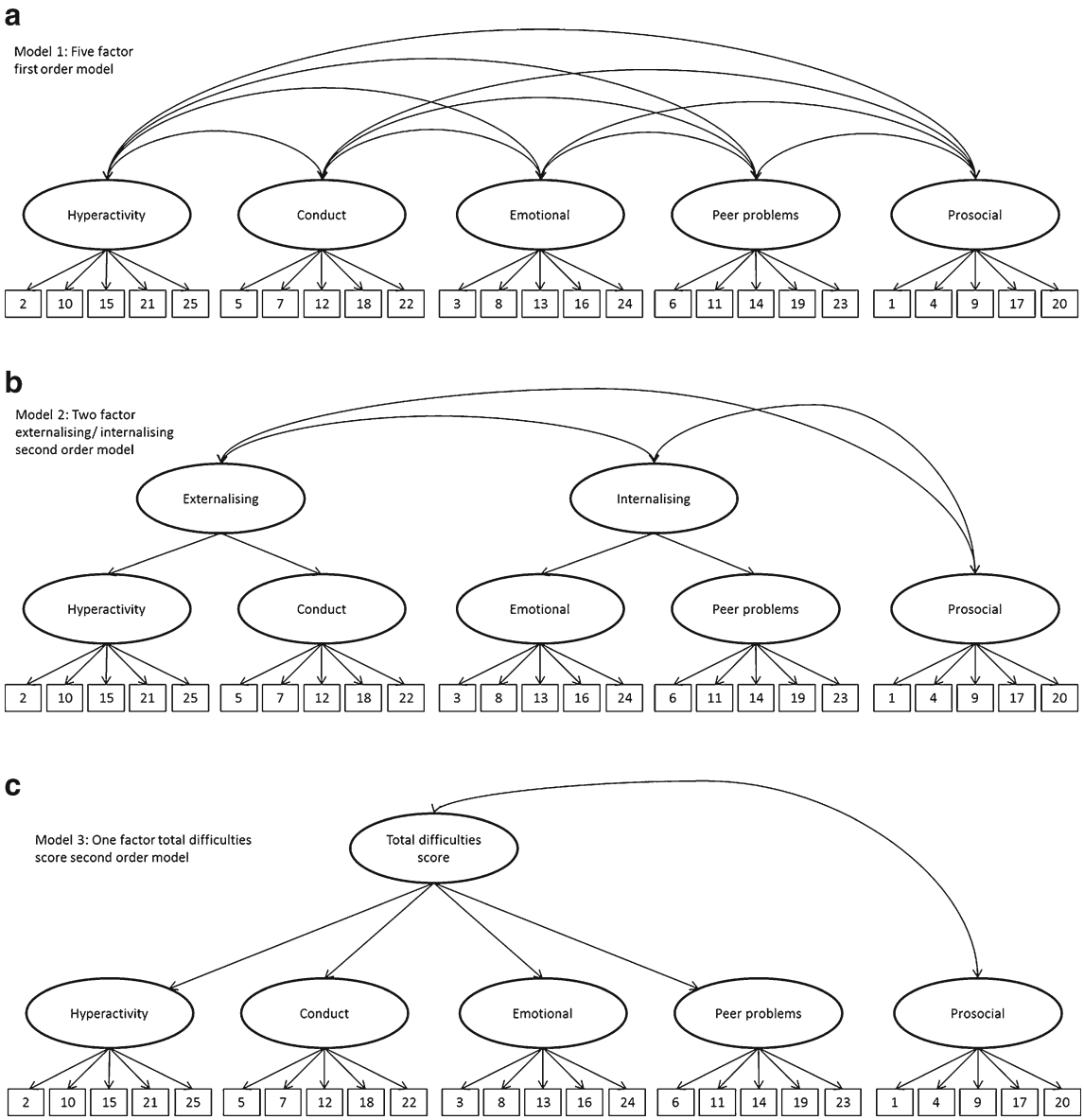

Fig. 1 Three theoretical models tested in CFA for each of the eight subgroups

2008; Christensen 2004; Elberling et al. 2010; Jacobsen et al. 2010; Nohr et al. 2006). The individual cohorts have been described in more detail elsewhere (Aarhus Birth Cohort 2008; Christensen 2004; Elberling et al. 2010;

Table 1 Characteristics of the birth cohorts providing SDQ data for the study

Cohort	Copenhagen Child Cohort	Danish National Birth Cohort	Danish National Institute of Social Research	Aarhus Birth Cohort
Acronym	CCC2000	DNBC	DNISR	ABC
Recruitment period	2000	1996–2002	1995	1990–1992
Study population: Eligible for the included follow-up	5,898	83,315[a]	5,233	8,244
Parent contribution of SDQ	3,349 (57 %)	48,544 (58 %)	4,971 (95 %)	6,751 (82 %)
Teacher contribution of SDQ	2,594 (44 %)	N/A[b]	N/A[b]	5,631 (68 %)
Age at SDQ screening	5	7	7	10–12

[a] As per October 2009; [b] *N/A* Not applicable

Olsen et al. 2001). Ethical approval was obtained for all of the studies.

Materials

The SDQ contains 25 questions asking about different positive and negative aspects of the child's behaviour. Responses are made on a three point Likert scale; *'not true','somewhat true'* and *'certainly true'*. Following the scoring recommendations, the items are divided into five subscales (*hyperactivity scale, emotional symptoms scale, conduct problem scale, peer problem scale and prosocial scale*) each comprising five items. The sum score of the first four subscales yields a *total difficulties score*. Parallel versions of the SDQ have been developed for parents, teachers and young persons (Goodman 1997; Goodman and Scott 1999).

Statistical Analyses

The method of Confirmatory Factor Analysis (CFA) was chosen as the appropriate means to test the three hypothesised models as it takes measurement error into account. All analyses were performed using the statistical package MPlus version 6.12. As the 25 items were rated on a non-redundant 3-point Likert scale and all items had skewed or indeed very skewed distributions, the data were treated categorically.

Previous research has found the weighted least square (WLS) method to be the superior estimator for CFA modelling of categorical data of exceptionally large samples sizes (Jöreskog and Sörbom 1996) and this estimator was applied for the two samples of younger boys and girls rated by parents (N=28,920 and 27,611 respectively). The weighted least square means and variance adjusted (WLSMV) on the other hand has been found to be superior with small to medium sample sizes (Brown 2006) and was initially applied for all analyses for the remaining six samples that varied in size between 1,272 and 3,322. The WLS estimator proved superior to the WLSMV within all samples and was therefore applied for all analyses for all samples throughout the study.

Model fits were evaluated by means of chi square test of model fit where 0 indicates a perfect fit, the Steiger-Lind root mean square error of approximation (RMSEA) where an RMSEA <0.08 indicates an acceptable model fit and <0.05 a good model fit, and Bentler comparative fit index (CFI) and Tucker-Lewis fit index (TFI), where CFI and TLI >0.90 signifies acceptable fits and >0.95 signifies good fits respectively (Schreiber et al. 2006). When certain parts of the model did not show acceptable fits, cross-loadings between specific indicators were allowed for on the basis of residual correlations and between factors and indicators based on modification indices. These modifications were only allowed for if they were considered to be theoretically meaningful.

Results

Missing Data

Kline suggests that less than 5 % of data missing on a single variable should be of little concern (Kline 2011). In the present study missing values were considered as missing at random (MAR); they constituted less than 0.05 % of all data and resulted in listwise deletion of cases. A further eleven cases were deleted due to lack of information on gender. The 71,840 cases were on these grounds reduced to 71,248.

Overall Model Fits: Factor Structure of the SDQ

Three different models were examined in the present study (Fig. 1). Model 1 was identical to Goodman's original factor structure with five hypothesised first-order factors (*hyperactivity/inattention, emotional, peer problems, conduct and prosocial)*. Model 2 added 2 second-order *internalising*/*externalising* factors to Model 1 and Model 3 added 1 second-order *total difficulties* factor to Model 1. All models were tested separately as a function of informants (parent and teachers), ages (younger and older) and gender (boys and girls), yielding a total of eight subgroups.

Initially the five separate scales (*hyperactivity, conduct, emotional, peer problems* and *prosocial*) were examined as five individual models with one factor and five indicators each in order to specify five separate well working models. This procedure was carried out for each of the eight subgroups separately. These were then aggregated to a full Model 1 for each sample. Having identified eight best working, theoretically justified models, a number of cross-loadings that improved the models for all of the eight subsamples were identified. This was done in order to identify one overall well working model for all subsamples. The following three cross-loadings between indicators were identified as yielding improved model fits across all samples: item 22 ("*steals from home school or elsewhere*") and item 18 ("*often lies or cheat*"); item 10 ("*constantly fidgeting or squirming*") and item 2 ("r*estless, overactive, cannot stay still for long*"); and item 20 ("*often volunteers to help others (parents, teachers, other children*)") with item 9 ("*helpful if someone is hurt, upset or feeling ill*"). These cross-loadings were not only permitted as they significantly improved model fits but also because they were considered theoretically meaningful. Items 22 and 18 are both concerned with delinquent behaviour, items 10 and 2 with

problems of keeping calm and sitting still and items 20 and 9 are both associated with helpful behaviour. Further, cross-loadings between the two positively worded items 21 ("*thinks things out before acting*") and item 14 ("*generally liked by other children*") were allowed to cross-load with the prosocial factor as this improved fit statistics significantly and was considered an appropriate means to capture response bias. Running Model 3 with these modifications resulted in non-convergent models with the implication that factor loadings could not be computed. This could indicate misspecifications in the model, or it could indicate that the model was overpowered because of the large sample sizes. As Model 3 was consistently found to have the poorest fits, these problems were not pursued further within the scope of this article. Thus for Model 3 only the raw, unadjusted model fits are presented.

Tables 2 and 3 present the initial measured, unadjusted model fits (in parentheses) along with the fits for the slightly modified models. The RMSEA values were considered good for all samples whereas the CFI and TLI were considered good for the teacher samples and acceptable for the parent samples. Inspection of the RMSEA, CFI and TLI revealed that Model 3 consistently had the poorest fits and further that the fits of Model 1 were generally somewhat better than the fits for Model 2. However, considering that Model 2 was the more parsimonious of the two models and considering that the differences of the fit statistics actually were minor, the fits for Model 1 and Model 2 were considered equally good. The two models could be compared statistically by means of the chi square difference test. However, since all the samples were large or extremely large, all yielded very large chi square values (Tables 2 and 3) and all chi square difference tests would in return be expected to prove highly significant. When such chi square difference tests were carried out they were indeed highly significant. This is because the data sets are so large and therefore overpowered, which means that even minor and trivial differences between the models will be found to be statistically highly significant. Because of this, the results of these analyses are not reported here. Another possible way to investigate whether there are true and meaningful differences between the models is by randomly selecting a number of smaller samples (e.g., N=250 or 500) drawn from the full cohort. If the differences remain in these smaller samples they can be considered nontrivial and important. This approach was carried out with N=250, 500 and 1000. However, most analyses resulted in non-identified models and results of these analyses are therefore not presented here.

Multi-group Analyses

In order to test for multi-group invariance the chi square contributions from each sample were used to carry out multi-group analyses for the modified Model 1 between parent and teacher raters and between boys and girls (Tables 4 and 5). As no information was available for the different age groups within the same samples, these analyses were not carried out. From the chi square values it appeared that the data fit Model 1 more convincingly for teachers than for parents. Possible reasons for this are described in more detail immediately below. It seems that higher factor loading and more explained total variance for individual items can explain at least part of the lower (and thus better) chi square values for teachers than for parents.

Table 2 SDQ parent Chi Square model fits, RMSEA, CFI and TLI for younger and older children and boys and girls separately. Fits with modifications (items 22–18, 10–2, 20–9 and the prosocial factor with positively worded items 21 and 14) are presented as are fits without modifications (in brackets)

Parent SDQ	Model	Chi Square	DF	RMSEA	CFI	TLI
Younger girls (N=27,611)	Model 1	7159 (10002)	260 (265)	0.031 (0.036)	0.893 (0.849)	0.877 (0.829)
	Model 2	7385 (10056)	263 (268)	0.031 (0.036)	0.890 (0.848)	0.874 (0.830)
	Model 3	(10688)	(270)	(0.037)	(0.839)	(0.821)
Younger Boys (N=28,920)	Model 1	8790 (12782)	260 (265)	0.034 (0.040)	0.906 (0.863)	0.892 (0.844)
	Model 2	9089 (12879)	263 (268)	0.034 (0.040)	0.903 (0.861)	0.889 (0.845)
	Model 3	(13642)	(270)	(0.041)	(0.853)	(0.837)
Older girls (N=3,237)	Model 1	1253 (1700)	260 (265)	0.034 (0.041)	0.934 (0.905)	0.924 (0.892)
	Model 2	1341 (1736)	263 (268)	0.036 (0.041)	0.929 (0.903)	0.919 (0.891)
	Model 3	(2123)	(270)	(0.046)	(0.911)	(0.901)
Older boys (N=3,322)	Model 1	1501 (2150)	260 (265)	0.038 (0.046)	0.938 (0.906)	0.928 (0.893)
	Model 2	1570 (2169)	263 (268)	0.039 (0.046)	0.935 (0.905)	0.925 (0.893)
	Model 3	(2265)	(270)	(0.047)	(0.900)	(0.889)

Table 3 SDQ teacher Chi Square model fits, RMSEA, CFI and TLI for younger and older children and boys and girls separately. Fits with modifications (items 22–18, 10–2, 20–9 and the prosocial factor with positively worded items 21 and 14) are presented as are fits without modifications (in brackets)

Teacher SDQ	Model	Chi Square	DF	RMSEA	CFI	TLI
Younger girls (*N*=1,291)	Model 1	1043 (1308)	260 (265)	0.048 (0.055)	0.955 (0.940)	0.948 (0.932)
	Model 2	1097 (1349)	263 (268)	0.050 (0.056)	0.952 (0.937)	0.945 (0.930)
	Model 3	(1458)	(270)	(0.058)	(0.931)	(0.924)
Younger boys (*N*=1,272)	Model 1	1100 (1502)	260 (265)	0.050 (0.061)	0.961 (0.943)	0.955 (0.935)
	Model 2	1132 (1542)	263 (268)	0.051 (0.061)	0.960 (0.941)	0.954 (0.934)
	Model 3	(1673)	265 (270)	(0.064)	(0.935)	(0.928)
Older girls (*N*=2,805)	Model 1	1491 (1903)	260 (265)	0.041 (0.047)	0.967 (0.957)	0.962 (0.951)
	Model 2	1513 (1935)	263 (268)	0.041 (0.047)	0.967 (0.956)	0.962 (0.951)
	Model 3	(2165)	(270)	(0.050)	(0.950)	(0.944)
Older boys (*N*=2,790)	Model 1	1903 (2515)	260 (265)	0.048 (0.055)	0.973 (0.963)	0.969 (0.958)
	Model 2	1953 (2535)	263 (268)	0.048 (0.055)	0.972 (0.963)	0.968 (0.958)
	Model 3	(2663)	(270)	(0.056)	(0.961)	(0.956)

Standardised Factor Loadings

One possible explanation for the differences in the multi-group analyses could be the observed differences in the standardised factor loadings; i.e., it is expected that the items (e.g., the five hyperactivity items) of an underlying factor (e.g., the hyperactivity scale) should show relatively high standardised loadings on that particular factor, but low loadings on other factors. Overall, higher loadings were found for the teacher samples compared to the parent samples (Table 6). Highest loadings were found for older children rated by teachers whereas lowest loadings were observed for younger children rated by their parents. No noteworthy differences were found between boys and girls. For all subsamples, the best parameter estimates were established for the hyperactivity scale indicating this to be psychometrically most satisfactory scale. Virtually all items on all scales were considered high for the teacher ratings and were all considered good. However, low standardised loadings were consistently found in most samples for the emotional item 3 ("*often complains of headaches, stomach-aches or sickness*"). It should be noted that the relatively low loadings of items 14 *("generally liked by other children")* and 21 *("thinks things out before acting")* are caused by their cross-loadings with the prosocial factor.

Explained Total Variances for the Observed Variables

Another plausible explanation for the differences reported in the multi-group analyses above are differences in the values of R^2 (Table 7). R^2 refers to the magnitude of proportion of variance for each observed variable that is accounted for by its related latent factor. Values of R^2 are computed by subtracting the square of the residual from 1. The values of R^2 should preferably be >0.50 indicating that at least 50 % of the total variance of that indicator has been explained by the model, with the remaining unexplained parts of the variance being attributable to other, residual factors. Values of $R^2<$ 0.50 are considered critically low since more than 50 % of the variance is then explained by factors other than the test item itself. The values of R^2 were consistently found to be much higher for teacher ratings than for parent ratings and also markedly higher for older children than for younger children. For older children with teacher raters, all R^2values explained more than 50 % of the total variance indicating that all items work well. By contrast, for younger children rated by their parents as many as 16 and 14 out of the 25 items (for girls and boys respectively) explained <0.50 of the total variance indicating severe problems with several test items for this age groups with parent raters. These

Table 4 chi square multi-group comparisons between parents and teachers. Chi Square contributions from each subsample is presented

	Parents	Teachers
Younger girls	1896 (*N*=1630)	1699 (*N*=1291)
Younger boys	1964 (*N*=1694)	1584 (*N*=1272)
Older girls	2263 (*N*=3237)	2153 (*N*=2805)
Older boys	2898 (*N*=3322)	2723 (*N*=2790)

Table 5 Multi-group comparisons between boys and girls. Chi Square contributions from each subsample is presented

	Boys	Girls
Parents younger children	9685 (*N*=28920)	8398 (*N*=27611)
Parents older children	1671 (*N*=3322)	1471 (*N*=3237)
Teachers younger children	1305 (*N*=1272)	1349 (*N*=1291)
Teachers older children	2120 (*N*=2790)	1772 (*N*=2805)

Table 6 Factor loadings for the separate parent and teacher samples for each of the indicators of the five latent variables (for the modified Model 1 that allows unique variance to correlate between factors and indicators)

	Items	Parent SDQ				Teacher SDQ			
		Younger girls	Younger boys	Older girls	Older boys	Younger girls	Younger boys	Older girls	Older boys
Hyperactivity/ Inattention	2	0.73	0.77	0.74	0.75	0.94	0.90	0.92	0.90
	10	0.69	0.72	0.71	0.66	0.93	0.88	0.85	0.86
	15	0.87	0.87	0.93	0.93	0.94	0.95	0.89	0.95
	21	0.49	0.51	0.53	0.69	0.64	0.53	0.61	0.60
	25	0.81	0.82	0.85	0.87	0.88	0.90	0.94	0.94
Emotional Problems	3	0.40	0.37	0.48	0.52	0.61	0.43	0.74	0.79
	8	0.61	0.62	0.70	0.74	0.81	0.82	0.82	0.80
	13	0.76	0.75	0.83	0.84	0.88	0.78	0.97	0.93
	16	0.66	0.67	0.75	0.82	0.80	0.84	0.87	0.86
	24	0.73	0.74	0.73	0.76	0.75	0.89	0.79	0.88
Conduct problems	5	0.57	0.62	0.67	0.71	0.87	0.87	0.90	0.87
	7	0.63	0.63	0.68	0.61	0.80	0.85	0.87	0.88
	12	0.77	0.82	0.85	0.81	0.95	0.91	0.96	0.93
	18	0.62	0.58	0.73	0.65	0.84	0.73	0.94	0.87
	22	0.48	0.46	0.56	0.53	0.87	0.66	0.87	0.84
Peer problems	6	0.60	0.69	0.65	0.73	0.84	0.89	0.86	0.85
	11	0.47	0.54	0.63	0.61	0.70	0.82	0.90	0.92
	14	0.58	0.61	0.68	0.74	0.62	0.55	0.65	0.58
	19	0.71	0.74	0.88	0.84	0.68	0.81	0.89	0.83
	23	0.67	0.75	0.81	0.85	0.88	0.88	0.84	0.86
Prosocial	1	0.84	0.87	0.83	0.91	0.92	0.95	0.96	0.96
	4	0.62	0.60	0.70	0.68	0.82	0.81	0.86	0.87
	9	0.58	0.59	0.63	0.66	0.78	0.82	0.81	0.87
	17	0.53	0.56	0.59	0.56	0.84	0.79	0.75	0.80
	20	0.46	0.45	0.54	0.52	0.53	0.65	0.71	0.73

marked differences of R^2 values between parent and teacher ratings can explain some of the differences in the multi-group analyses found above. Virtually no differences were observed in R^2 values between boys and girls. The value of $R^2$for item 3 *("often complaints of headaches…")* was the lowest for virtually all subsamples. For parent raters and younger children the item was consistently and critically low (e.g., for parents rating younger boys: 0.367^2=13.5 % of the total variance, leaving 86.5 % unexplained). Neither allowing the item to load on to other items, or factors nor removing the item altogether, increased either the general fits of the models, or the total variance explained by that item.

Reliability Measures

In order to evaluate the internal consistency of the individual scales, i.e., to what degree the scores are free from random measurement error, composite reliability (CR) and average variance extracted (AVE) were calculated. Although Cronbach's Alpha is the most commonly used measure of reliability in the literature, it is not reported here as it is a conservative measure of reliability which assumes that all items contribute equally to the reliability, i.e., it estimates how the full scale works rather than taking account of the variance and measurement error of the individual items. The AVE and CR on the other hand are reported here as they take complexity into account and do not assume that all items add equally to the reliability of the factor in question. CR is specifically concerned with the composite of the items taking into account the standardised loadings and the measurement errors of each of them. If CR >0.70 then satisfactory scale reliability is typically considered to have been established. AVE on the other hand is a measure that indicated how much variance is, on average, explained. If an item is overall poor for its scale it will result in a low AVE (<0.50) (Fornell and Larcker 1981). It appears from Table 8 that all CR's were above 0.7 indicating good scale reliability for all scales for all subsamples. It should be noted, however, that the lowest values of CR were found for younger children with parent raters and highest values were found for older

Table 7 values of R^2 for the separate parent and teacher samples for each of the indicators of the five latent variables (for the modified Model 1 that allows unique variance to correlate between factors and indicators)

	Items	Parent SDQ				Teacher SDQ			
		Younger girls	Younger boys	Older girls	Older boys	Younger girls	Younger boys	Older girls	Older boys
Hyperactivity/ Inattention	2	0.53	0.59	0.54	0.56	0.89	0.80	0.85	0.81
	10	0.47	0.53	0.50	0.43	0.86	0.77	0.72	0.74
	15	0.76	0.76	0.86	0.87	0.88	0.90	0.80	0.90
	21	0.44	0.49	0.52	0.56	0.75	0.68	0.74	0.76
	25	0.65	0.67	0.73	0.76	0.77	0.80	0.89	0.88
Emotional problems	3	0.16	0.14	0.23	0.27	0.37	0.18	0.55	0.62
	8	0.37	0.39	0.49	0.55	0.66	0.66	0.68	0.64
	13	0.58	0.56	0.69	0.70	0.77	0.61	0.94	0.86
	16	0.43	0.46	0.56	0.67	0.64	0.71	0.75	0.74
	24	0.53	0.55	0.54	0.58	0.57	0.79	0.62	0.77
Conduct problems	5	0.32	0.39	0.45	0.51	0.75	0.76	0.81	0.76
	7	0.40	0.40	0.46	0.37	0.63	0.72	0.76	0.78
	12	0.59	0.67	0.73	0.66	0.90	0.83	0.92	0.87
	18	0.39	0.33	0.53	0.42	0.71	0.53	0.89	0.75
	22	0.23	0.21	0.32	0.28	0.75	0.43	0.76	0.70
Peer problems	6	0.36	0.47	0.42	0.53	0.71	0.79	0.73	0.72
	11	0.22	0.29	0.40	0.37	0.50	0.68	0.80	0.85
	14	0.54	0.64	0.73	0.76	0.86	0.87	0.91	0.90
	19	0.51	0.54	0.78	0.71	0.47	0.65	0.79	0.69
	23	0.45	0.56	0.66	0.72	0.78	0.76	0.71	0.74
Prosocial	1	0.71	0.76	0.69	0.83	0.85	0.91	0.93	0.91
	4	0.38	0.36	0.49	0.46	0.68	0.66	0.73	0.76
	9	0.33	0.35	0.40	0.43	0.62	0.67	0.65	0.75
	17	0.28	0.31	0.35	0.31	0.70	0.63	0.56	0.64
	20	0.21	0.20	0.30	0.27	0.28	0.43	0.51	0.54

Table 8 Composite Reliability (CR) and Average Variance Extracted (AVE) for the separate parent and teacher subsamples

	Reliability	SDQ parents				SDQ teachers			
		Younger girls	Younger boys	Older girls	Older boys	Younger girls	Younger boys	Older girls	Older boys
Hyperactivity/ inattention	CR	0.86	0.82	0.88	0.89	0.96	0.94	0.95	0.95
	AVE	0.55	0.48	0.61	0.62	0.82	0.77	0.78	0.80
Emotional problems	CR	0.77	0.77	0.83	0.86	0.88	0.87	0.92	0.93
	AVE	0.41	0.42	0.50	0.55	0.60	0.59	0.71	0.72
Conduct problems	CR	0.75	0.76	0.83	0.80	0.94	0.90	0.96	0.94
	AVE	0.39	0.40	0.50	0.45	0.75	0.65	0.82	0.77
Peer problems	CR	0.76	0.82	0.87	0.88	0.89	0.93	0.94	0.94
	AVE	0.39	0.47	0.58	0.60	0.63	0.72	0.77	0.75
Prosocial	CR	0.75	0.76	0.79	0.80	0.89	0.90	0.91	0.93
	AVE	0.38	0.40	0.44	0.46	0.62	0.66	0.68	0.72
Externalising	CR	0.87	0.88	0.91	0.90	0.94	0.90	0.95	0.95
	AVE	0.78	0.79	0.83	0.82	0.89	0.82	0.90	0.91
Internalising	CR	0.79	0.82	0.85	0.88	0.88	0.88	0.91	0.89
	AVE	0.66	0.69	0.73	0.78	0.78	0.79	0.84	0.81
Total	CR	0.88	0.87	0.87	0.92	0.94	0.91	0.96	0.95
	AVE	0.64	0.64	0.63	0.76	0.78	0.73	0.85	0.83

children with teacher raters, indicating that the individual scales work better in the latter situation. No substantial differences were found between boys and girls. From the sizes of AVE it appears that all factors work well for older children rated by teachers and also that no items from the hyperactivity/inattention subscale are problematic for any of the subsamples. Single items on the emotional, conduct, peer and prosocial scales, on the other hand, do create problems for these scales for younger children rated by parents, resulting in poor values of AVE. This is, however, not surprising since 14 items and 16 items out of 25 explained <0.50 of the total variance for these samples of boys and girls respectively.

Discussion

The aim of the present study was to examine how well 71,248 SDQ ratings, divided into eight subgroups, fit three theoretically based models by means of confirmatory factor analysis. It was concluded that Model 1including five latent first order factors and Model 2 including a further two *internalising*/*externalising* second order factors both have good fits and work equally well. Model 3, which included one *total difficulties* second order factor was throughout all samples found to be less satisfactory than Model 1 and Model 2. Also, data from teachers seem to fit the models better than data from parents, and from older children better than for younger ones. No differences were found between boys and girls. Although Model 1 and Model 2 overall are well working several of the findings call for closer inspection.

Firstly, it appears that both Model 1 and Model 2 show good overall fit statistics in the present study of low risk epidemiological samples. This finding is somewhat in contrast to Goodman et al. (2010) who concluded that the broader internalising and externalising SDQ subscales of Model 2 are superior in low risk epidemiological samples. Contrary to the present study, however, Goodman et al. did not subdivide their sample on the basis of gender or age. Considering then that the present study did so and did find rather large intergroup differences on the basis of age and rater (but not on the basis of gender), this may partly explain these somewhat contradictory findings. The findings of the present study suggest that different models can be advantageously examined by subdividing the sample on the basis of age and rater (but not necessarily on gender) leading to better and more accurate model fits. The sample used in the study by Goodman et al. included 5–16-year-old children and this large age span may have masked potential differences between subgroups. Another potential explanation for this discrepancy in results is that the differences are genuine and are caused by cultural differences. Compared to Goodman's British cultural setting, Denmark is probably more homogenous in terms of access to the education and health care systems—services that are all tax-financed and free of charge for the citizens. Examining the influence of such cultural and societal differences on the factor structure of the SDQ remains to be carried out, but it certainly would be both an interesting and highly relevant study for future research considered the global and widespread use of the SDQ.

Secondly, all models significantly benefitted from minor model modifications, i.e., allowing the two positively worded items 14 and 21 to cross-load with the prosocial factor and allowing cross-loadings between items 22–18, 10–2 and 20–9. These modifications represent systematic, rather than random, measurement errors in item responses and they may derive from characteristics which are specific either to the test items or to the respondents (Byrne 2011). In other words, reversed items 14 and 21 not only relate to their respective factors (*peer problems* and *hyperactivity*) but they also reflect response bias and some underlying prosocial behaviour. It is recommended that that the above model modifications should be applied for future research purposes. This is, however, not feasible in clinical settings and there it is instead recommended that sum scores be retained, as also originally recommended by Goodman.

Thirdly, one of the major advantages of structural equation modeling is that it provides a comprehensive means for assessing and modifying theoretical models. The findings presented in the present paper suggest a cautious future use of positively worded, reversed items in questionnaires of this type, as this may contaminate the factor structure of the questionnaire. The present study tested for, but did not find, support for a positive construct factor (the results were not presented here). However, there are still many and very good reasons to include positively worded items in a questionnaire of this type. Firstly, as noted by Goodman, because it enhances acceptability of the questionnaire on the part of the rater, especially so in the general population (Goodman 1994). Secondly, because it expands the description of the mental health functioning of the child by including non-pathological traits. By adding assessment of mental health strengths, the questionnaire informs about possible protective or resilience factors, which might be of particular importance in the investigation of developmental psychopathology.

Fourthly, item 3 *"often complains of headaches, stomach-aches or sickness"* repeatedly showed poor factor loadings and explained critically little of the total variance throughout most analyses. Neither removing the item, nor allowing the item to cross-load with other items or scales improved the model. However, it was retained in the model as it did contribute significantly to the overall model fits. There may be several reasons for this item fitting the SDQ so poorly: 1. from a closer inspection of the wording of the item, it appears that it is actually the only one of the 25 items that relies on some sort of self-report on the part of the child. The remaining 24 items solely rely on evaluation on part of

the rater. 2. The item appears to be *state* dependent reflecting the state of the child at a particular moment in time, whereas the remaining 24 items appear to reflect traits i.e., relatively time-stable individual characteristics. In other words, the item may represent an unspecific marker of impact, probably expressed by age appropriate somatic symptoms, rather than as a direct psychopathological trait or symptom. 3. very little of item 3's total variance is explained. In other words, when children complain of headaches, one cannot be certain that they actually have a headache. Instead, it may indicate that they experience other sorts of unspecified problems.

Finally, the questionnaire was found to be superior for teacher compared to parent raters and for older children compared to younger ones. These differences were found between the different subsamples on all levels of analyses, namely on an overall model level, a factor level and an item level and they point to the importance of running at least age and rater specific analyses in future work with the SDQ.

Limitations and Future Work

A limitation of the present study is the lack of access to a high risk sample. It is not known whether one model would prove superior to the other within such a setting, as was concluded in the study by Goodman et al. (2010). Future studies should replicate the analyses of the present study using high risk, clinical samples, in order to investigate whether the present findings hold true across such groups. The participation rates in some of the published samples are rather low and this could potentially have had an effect on the results. The substantial size of the sample has allowed for very specific comparisons of item functioning across the different samples of ages, raters and gender. Such analyses were beyond the scope of the present article but will be a highly relevant focus for future studies.

Acknowledgments I would like to thank associate professor Thomas William Teasdale, PhD DM Sci. for commenting on an earlier draft of this manuscript and Jan Ivanouw, M.Sc. PhD for psychometrics advice.

Funding The study was financially supported by Institute of Psychology, University of Copenhagen; the Lundbeck Foundation; The Carl J Beckers Foundation; Børne- og Ungdomspsykiatrisk Selskab i Danmark; Direktør Jacob Madsen og Hustru Olga Madsens Foundation; The A.P. Møller Foundation for the Advancement of Medical Science; The Dagmar Marshall Foundation; Aase og Ejnar Danielsens Foundation; The Ludvig and Sara Elsass Foundation.

References

Aarhus Birth Cohort. (2008). www.aarhus-born.dk.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders*. Washington DC: DSM-IV.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. (2011). *Structural equation modeling with MPlus*. Routledge.

Christensen, E. (2004). *7 års børneliv. Velfærd, sundhed og trivsel hos børn født i 1995 (Rep. No. 04:13)*. København: Socialforskningsinstituttet.

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the strengths and difficulties questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry, 43*, 1159–1167.

Elberling, H., Linneberg, A., Olsen, E. M., Goodman, R., & Skovgaard, A. M. (2010). The prevalence of SDQ-measured mental health problems at age 5–7 years and identification of predictors from birth to preschool age in a Danish birth cohort: The Copenhagen Child Cohort 2000. *European Child & Adolescent Psychiatry, 19*, 725–735.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Goodman, R. (1994). A modified version of the Rutter parent questionnaire including extra items on children's strengths: a research note. *Journal of Child Psychology and Psychiatry, 35*, 1483–1494.

Goodman, R. (1997). The strengths and difficulties questionnaire: a research note. *Journal of Child Psychology and Psychiatry, 38*, 581–586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1337–1345.

Goodman, R., & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the child behavior checklist: is small beautiful? *Journal of Abnormal Child Psychology, 27*, 17–24.

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (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. *Journal of Abnormal Child Psychology, 38*, 1179–1191.

Jacobsen, T. N., Nohr, E. A., & Frydenberg, M. (2010). Selection by socioeconomic factors into the Danish National Birth Cohort. *European Journal of Epidemiology, 25*, 349–355.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference gouide*. Chicago: Scientific Software International.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. (Third ed.) Guilford.

Koskelainen, M., Sourander, A., & Kaljonen, A. (2000). The strengths and difficulties questionnaire among Finnish school-aged children and adolescents. *European Child & Adolescent Psychiatry, 9*, 277–284.

Niclasen, J., Teasdale, T. W., Andersen, A. M., Skovgaard, A. M., Elberling, H., & Obel, C. (2012). Psychometric properties of the Danish strength and difficulties questionnaire: The SDQ assessed for more than 70,000 raters in four different cohorts. *PLoS One, 7*, e32025.

Nohr, E. A., Frydenberg, M., Henriksen, T. B., & Olsen, J. (2006). Does low participation in cohort studies induce bias? *Epidemiology, 17*, 413–418.

Olsen, J., Melbye, M., Olsen, S. F., Sorensen, T. I., Aaby, P., Andersen, A. M., et al. (2001). The Danish national birth cohort—its background, structure and aim. *Scandinavian Journal of Public Health, 29*, 300–307.

Palmieri, P. A., & Smith, G. C. (2007). Examining the structural validity of the strengths and difficulties questionnaire (SDQ) in a U.S. sample of custodial grandmothers. *Psychological Assessment, 19*, 189–198.

Sanne, B., Torsheim, T., Heiervang, E., & Stormark, K. M. (2009). The strengths and difficulties questionnaire in the Bergen child study:

a conceptually and methodically motivated structural analysis. *Psychological Assessment, 21*, 352–364.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *The Journal of Educational Research, 99*, 323–337.

Van, R. B., Veenstra, M., & Clench-Aas, J. (2008). Construct validity of the five-factor strengths and difficulties questionnaire (SDQ) in pre-, early, and late adolescence. *Journal of Child Psychology and Psychiatry, 49*, 1304–1312.

World Health Organisation. (1993). The ICD-10 classification of mental and behavioural disorders diagnostic criteria for research.