ACTA UNIVERSITATIS CAROLINAE
KINANTHROPOLOGICA Vol. 52, 1 – 2016 Pag. 15–37

UNIVERSITY OF YORK, UNITED KINGDOM,
DEPARTMENT OF HEALTH SCIENCES[1]
CHARLES UNIVERSITY IN PRAGUE, CZECH REPUBLIC,
DEPARTMENT OF KINANTHROPOLOGY[2]
UNIVERSITY OF CAMBRIDGE, UNITED KINGDOM,
DEPARTMENT OF PSYCHIATRY[3]
BRADFORD INSTITUTE FOR HEALTH RESEARCH, UNITED KINGDOM[4]
UNIVERSITY OF DUNDEE, UNITED KINGDOM SCHOOL OF NURSING AND
HEALTH SCIENCES, DUNDEE CENTRE FOR HEALTH
AND RELATED RESEARCH[5]

THE PSYCHOMETRIC PROPERTIES OF THE STRENGTHS AND DIFFICULTIES QUESTIONNAIRE IN A MULTI-ETHNIC SAMPLE OF YOUNG CHILDREN

JAN ŠTOCHL[1,2,3], STEPHANIE L. PRADY[1], ELIZABETH C. ANDREWS[4],
KATE E. PICKETT[1,4], TIM CROUDACE[5]

ABSTRACT

Objectives: Developed countries are becoming more multi-ethnic, with consequent problems of maintaining invariance of health questionnaires. We aimed to explore commonalities in the structure of the Strengths and Difficulties Questionnaire (SDQ) completed for young children in a multi-ethnic English cohort while examining potential method effects and misfitting items. The secondary aim was to demonstrate the usefulness of bifactor modelling and exploratory structural equation modelling (ESEM) for kinanthropological research.

Methods: We used SDQ data from 3,290 children enrolled in the Born in Bradford cohort, completed by parents (usually the mother) at child age 3 and 4 and teachers at age 5. The factor structure for 11 potential configurations was assessed in each age group using confirmatory factor analysis. ESEM was used to assess misfitting items under the best fitting configuration.

Results: The best fitting configuration was a bifactor model of the 2 broader scales and a methods factor, using the 20 difficulties items. Generally, factor loadings increased between age 3 and age 5. Several items contributed to misfit.

Conclusions: There was less support for the robustness and hypothesised structure of the SDQ in this sample. Bifactor scores that account for measurement error could be useful if carefully applied in epidemiological and kinanthropological studies in multi-ethnic and/or younger age samples.

Keywords: Strength and Difficulties Questionnaire; factor analysis; exploratory structural equation modeling

DOI: 10.14712/23366052.2016.2

INTRODUCTION

The Strengths and Difficulties Questionnaire (SDQ) is a 25-item instrument used to screen for behavioural problems and is widely used as an indicator of psychopathological risk in general population surveys of children (Goodman, 1997). Its focus makes it attractive also for kinanthropological research, especially to explore associations between psychopathology signs in childhood and levels of physical activity either at young age (Hamer et al., 2009) or adolescence (Sagatun et al., 2007; Ussher et al., 2007). The SDQ has also been successfully used in a more specific kinanthropological context, for example, to measure associations between emotional problems and physical activity (Wiles et al., 2008) or to explore effects of television viewing on psychological health with physical activity as a covariate (Page et al., 2010). Finally, the SDQ was recently applied in the context of physiopathology such as low back pain (Watson et al., 2003), asthma (Glazebrook et al., 2006), cerebral palsy (Majnemer et al., 2008), sleep problems (Nixon et al., 2008), and other physical or neurological disabilities (Law et al., 2007).

There are several versions of the SDQ intended to be used to assess children age 4–17 which are formatted and worded to be completed by parents and by teachers (termed the SDQ^{4-17}), and a self-rated version for children age 11–17. More recently, the SDQ has been applied in samples of younger children. Early-years (child age 2–4) versions have been developed for teachers and parents (the SDQ^{2-4}). In these versions the developers have altered the wording of three questions.

For all SDQ versions, each item requires a response on scales that display three response options (that the described behaviour is Not True, Somewhat True or Certainly True), so that ratings can characterise three levels of strengths or difficulties. Twenty of the 25 questions describe aspects of problem behaviour. These problem behaviour items can be grouped into four (5-item) subscales representing 1) emotional, 2) peer, 3) behavioural and 4) conduct domains. Morbidity (problems) are indicated by higher scores. The four subscales can also be clustered into two 10-item broader scales with the emotional and peer subscales indicating internalising problems, and the hyperactivity and conduct subscale indicating externalising problems (Figure 1). The broader subscales have been suggested for use in community, rather than clinical, samples (Goodman et al., 2010a). Together, all 20 are used to provide a summary score for 'total difficulties' (Figure 1). Five of these 20 questions are positively worded, but with scoring reversed to reflect difficulties. These are shaded in Figure 1 and labelled 'methods'. The five remaining questions make up the prosocial subscale. Prosocial items are positively worded questions designed to indicate strengths rather than difficulties. This subscale is not usually included when assessing difficulties.

The risk of psychopathology can be estimated in a variety of ways using thresholds based on the summary score of total difficulties (20 questions), the two broader scales (2×10 questions), or the four subscales (4×5 questions). Mean score differences between groups for the broader scales or total difficulties score can also be estimated. Internal consistency for the subscales is satisfactory and higher for teacher rated scoring (weighted mean subscale score range 0.63 to 0.83 for $N = 26$ studies) than parent rated (range 0.53 to 0.76) with higher internal consistency for the total difficulties score (teacher rated 0.82, parent rated 0.80) (Stone et al., 2010). Consistently, and regardless

of rater, the peer subscale appears to be the least reliable. Test-retest reliability is also stronger for teacher rated scores (weighted mean correlation subscale score range 0.72 to 0.85 for N = 6 studies, total difficulties 0.84) than parent rated (subscale range 0.65 to 0.71, total difficulties 0.76) (Stone et al., 2010).

The SDQ can also be used to predict disorder (unlikely, possible, probable) from studies that have ratings from at least two different informants, for example from parents and teachers (Goodman et al., 2000b). The predictive algorithm utilises information from the multiple ratings on the subscales that indicate hyperactivity problems, conduct problems and emotional behaviour and the questions at the end of the SDQ that form an impact statement about the effect of the child's difficulties (Goodman et al., 2000b). The sensitivity of the algorithm to detect any psychiatric disorder via multi-informants in an English speaking sample of 5–15 year olds was 63.3%, with specific disorder sensitivity ranging from 50.1% (any anxiety disorder) to 86.1% (any hypokinetic disorder) (Goodman et al., 2000a). There was a high number of false positives in this community sample; 47.3% of children with a 'disorder probable' SDQ prediction did not have a psychiatric disorder, and misclassification errors are twice as likely using data from just one informant.

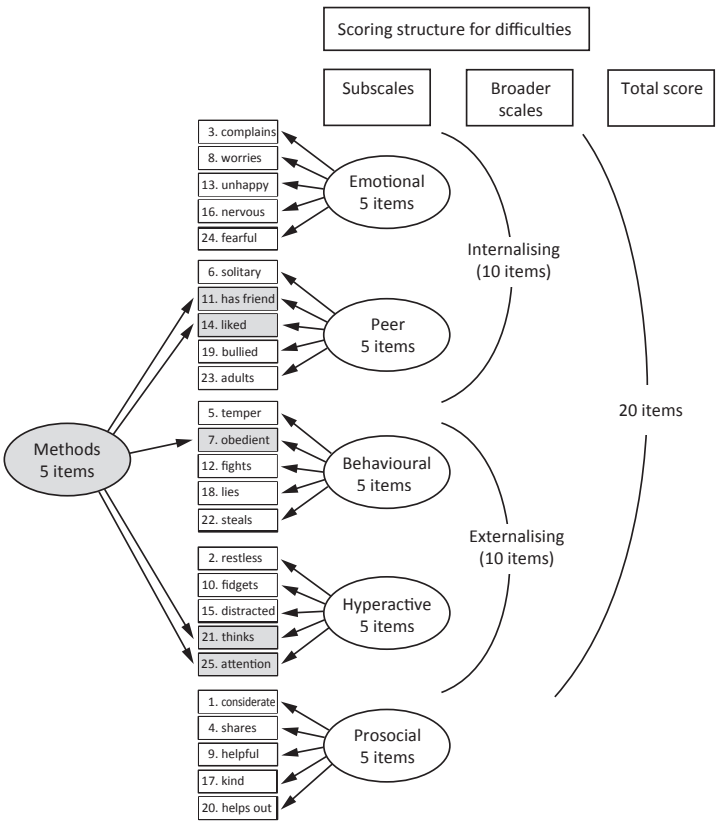


Figure 1. Component items of the Strengths and Difficulties Questionnaire and scoring structure

The SDQ$^{4–17}$ is the original scale and numerous studies have aimed to explore or confirm its hypothesised structure. A 2010 review found eight studies reporting confirmatory factor analyses (CFA), with some, but not universal, support found for structures represented by all five subscales, and the two broader scales plus the prosocial subscale (Stone et al., 2010). Some studies have reported an effect of the five positively worded difficulties questions (methods items) on the structure (McCrory & Layte, 2012; van de Looij-Jansen et al., 2011).

The SDQ$^{2–4}$ has only more recently been suggested as a behavioural questionnaire for initial risk assessment i.e. a screening instrument, and fewer modelling studies have been conducted in samples within this younger age range. A recently reported CFA of the parent administered SDQ at age 3 in a UK-representative sample found more support for a five than a three or one subscale structure (Croft et al., 2015). In contrast, CFA studies in Spanish and Dutch samples reported less than satisfactory baseline fit for the five subscales, a unidimensional structure of all 25-items, and a second order structure of the five subscales within the two broad subscales (Ezpeleta et al., 2013; Theunissen et al., 2013). As with analyses of the SDQ$^{4–17}$, the tested structures included items from the prosocial subscale even though this subscale is not typically included when scoring the instrument. Croft et al. (2015) studied invariance over time (age 3, 5 and 7), finding support for strong invariance for the conduct and hyperactivity subscales and metric invariance for the peer and emotional subscales. Presently – in this age range – there are few other correlation and descriptive studies, (e.g. Fuchs et al., 2013; Petermann et al., 2010).

The SDQ has, at the time of writing, been translated into 79 languages, and there are many studies reporting validity of translated versions. A potential problem of validity arises when several sub-populations, such as those indicated by different language, or ethnic group, are present in a sample, or when several samples with some language or cultural variation need to be compared. Robustness of inferences in populations with such sub-population variation are typically empirically investigated using differential item functioning (DIF) methodology (Gregorich, 2006). Findings from two cross national studies including children from a range of ages suggest that the SDQ cannot be assumed to have the same factorial structure or relationship with diagnoses of disorder across countries (Goodman et al., 2012; Stevanovic et al., 2014). Some studies have assessed potential measurement differences arising from cultural and language variation within countries in younger age samples. Two CFA studies in samples that included children age 4 or 5 have examined DIF by ethnicity using both teacher and parent rated SDQ$^{4–17}$. A UK study with only English versions of the SDQ found invariance between White and Indian sub-samples (Goodman et al., 2010b). A North American study also reported invariance between American English and Spanish language samples, however the baseline fit for each group, required to test invariance, were only marginally adequate (Hill & Hughes, 2007). A third study employing principal components analysis using data from younger children (mean age 5.3) across five ethnic groups in the Netherlands found evidence of DIF (Mieloo et al., 2014).

The implications for lack of invariance are clear; groups within samples cannot be compared with any accuracy and attempts to do so could lead to spurious conclusions regarding the difficulties of a particular ethnic group. Taking a forward view, as developed countries become more multi-ethnic there will be a growing need to validate instruments across increasingly diverse samples. Studies will include many ethnic groups,

with different sample sizes. This presents some difficulties. Foremost, any classification of a person into an 'ethnic group', is an artificially constructed analytical grouping of convenience that can mask important social and experienced variation (Nazroo, 1998). Factors such as cultural, racial and ethnic identity might vary in their interaction during the multidimensional process of acculturation, meaning that within-group variation could be greater than differences between-group, and these can change over time (Bhugra, 2005). Populations are dynamic and it may not be reasonable to assume homogeneity in the lived experience of persons of mixed and multi-ethnic heritage in particular; the fastest growing UK minority groups. Together with the technical difficulties that arise when assessing DIF in small sample sizes fragmented by attempts to classify 'homogeneous' groups, it is possible that a fresh approach towards multi-ethnicity in epidemiological studies is needed. If the relevant validity question were to be rephrased *what is common across a diverse sample, or between diverse samples?* Then it may be possible to find a valid structure that measures important dimensions of children's behaviour whilst also minimising measurement error and retaining already established validity claims.

In summary, despite the popularity of the SDQ, there is little work as yet describing the psychometric structure of the SDQ by the age of the child, and less research overall in multi-ethnic, younger age samples. The Born in Bradford (BiB) study administered the SDQ in three separate sub-studies at child age 3, 4, and 5, thus providing an ideal environment in which to examine commonalities in structure across variation in informants and age. We aimed to revisit the construct validity of the SDQ in light of recent research interest in examining risk of psychopathology and its association to physical activity in young children. Our secondary aim, though implicit, was to introduce bifactor modelling and exploratory structural equation modelling as well as demonstrate its usefulness for assessment of structural hypotheses and model misfit to wider kinanthropological community.

METHODS

Sample

In this analysis we used data collected in several Born in Bradford (BiB) sub-studies. Bradford is a city of around 500,000 inhabitants in the North of England with high levels of socio-economic deprivation and ethnic diversity, and BiB was set up to examine the impact of environmental, psychological and genetic factors on maternal and child health (Raynor and Born in Bradford Collaborative Group, 2008; Wright et al., 2013). Between 2007–2010 more than 12,000 women were recruited during pregnancy, of which 45% identified themselves as being of Pakistani origin, 39% White British and the remaining 6% of different and varying ethnicity. Three sub-studies have collected early SDQ data on BiB children:

1. At age 3, the Parent SDQ$^{2-4}$ was completed by the parent using computer-assisted personal interviewing (CAPI) for a study on childhood obesity (N = 1,217)
2. At age 4, parents filled in the Parent SDQ$^{4-17}$ via CAPI for a study on asthma (N = 1,711)
3. At age 5 the child's teacher filled in the Teacher SDQ$^{4-17}$ between March to July in 2013 and 2014 using a paper version of the SDQ (N = 2,365)

In all, there was at least one SDQ rating 3,920 children. Nominally, all three samples were broadly representative of children in the BiB study.

Scoring

Negatively worded difficulties items and the items in the prosocial scale were scored in each response category of Not True = 0, Sometimes True = 1 or Certainly True = 2. Scoring was reflected for responses to the five positively worded difficulties items.

Psychometric analysis

a) Confirmatory factor analysis

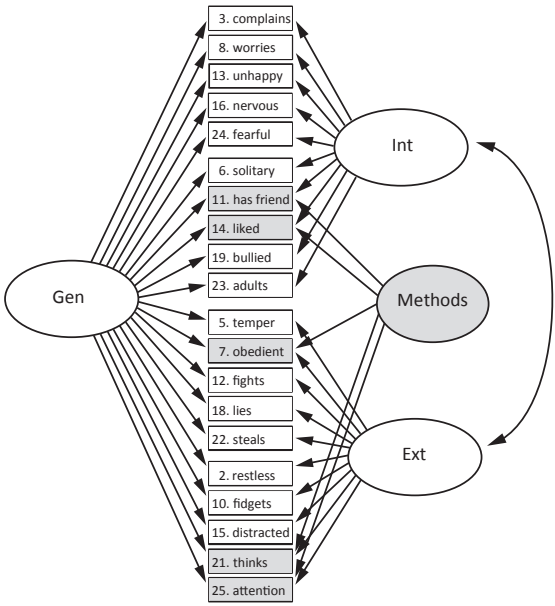
For each sample we tested 11 plausible structures that are typically examined (see Figure 2) in the SDQ using CFA under Full Information Maximum Likelihood (FIML; Enders and Bandalos, 2001) estimation[1]. In all three samples we noted fit indices of each configuration (Akaike Information Criterion (AIC; Akaike, 1973), Bayesian Information Criterion (BIC; Schwarz, 1978)). These relative fit indices allow for comparison of competing models (based on the same data but with varying numbers of parameters) since they penalize more complex models. The model with the lowest AIC and BIC is preferred.

Absolute model fit was assessed by means of Comparative Fit Index (CFI; Bentler, 1990), Root Mean Square Error of Approximation (RMSEA; Steiger and Lind, 1980) and Weighted Root Mean Square Residual (WRMR; Muthén and Muthén, 1998–2016a; Yu, 2002). CFI values larger than 0.95, RMSEA values lower than 0.06 and WRMR values around 1 are considered indicate model fit to data. More detailed recommendations on cut-off values for these fit indices can be found in Hu and Bentler (1999). To obtain this latter set of indices for assessment of absolute model fit, we re-estimated all models using mean and variance adjusted Weighted Least Squares (WLSMV; Muthén, 1993) as they are not provided when FIML is used.

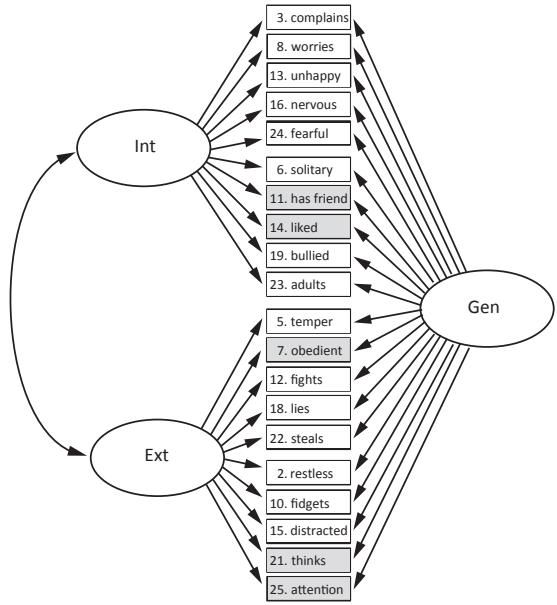
For further assessment of model misfit, we used an Exploratory Structural Equation Modeling (ESEM) approach on the best-fitting configuration. Mplus version 7.3 (Muthén and Muthén, 1998–2016b) was used for all analyses.

[1] A sandwich estimator was used to account for clustering of pupils by teacher in the age 5 sample.

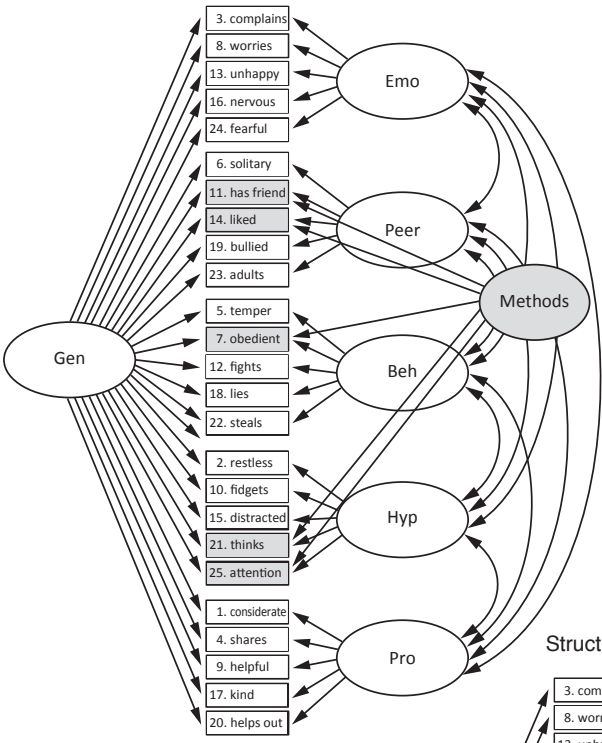
Structure 1



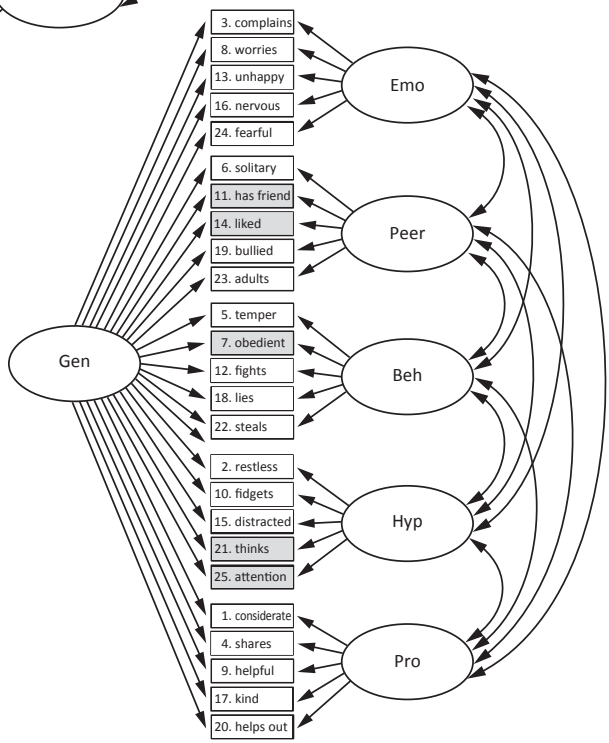
Structure 2



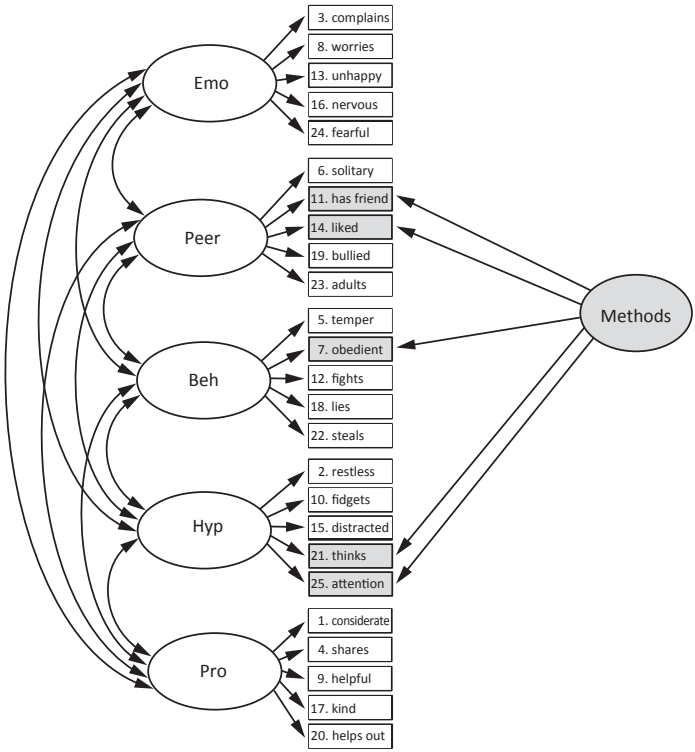
Structure 3



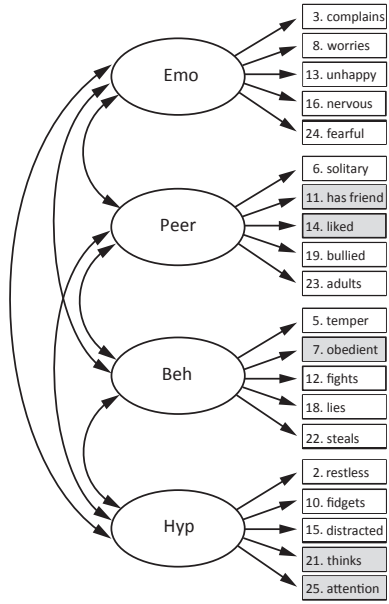
Structure 4



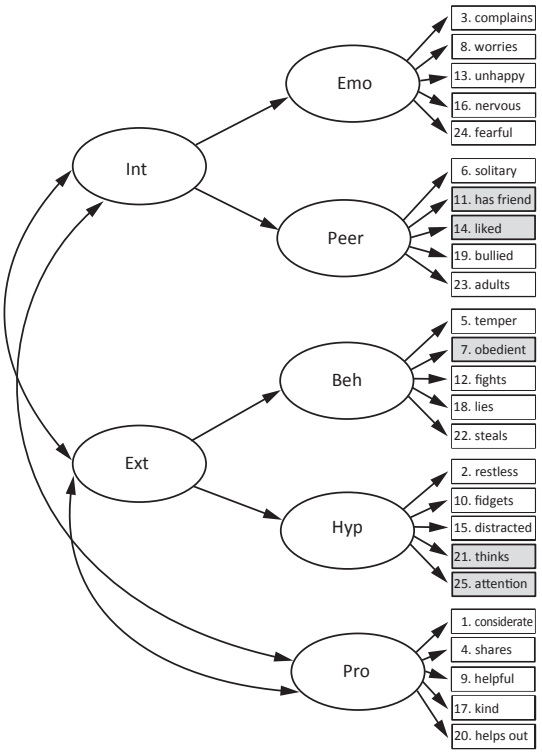
Structure 5



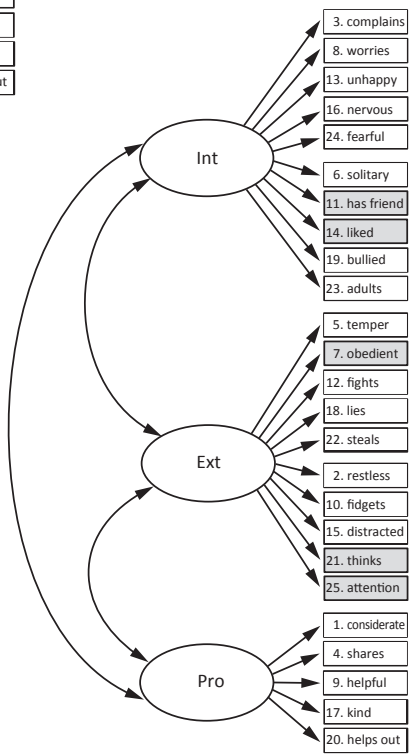
Structure 6



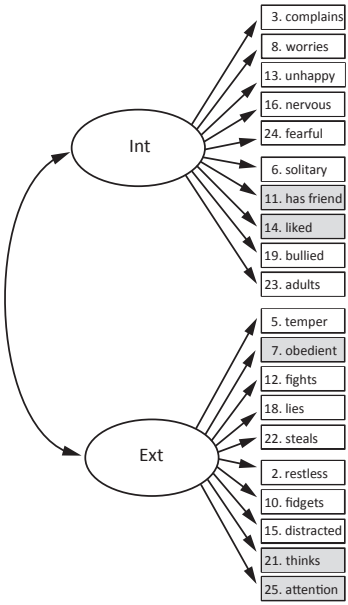
Structure 7



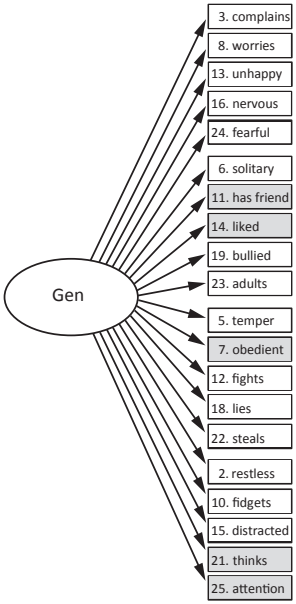
Structure 8



Structure 9



Structure 10



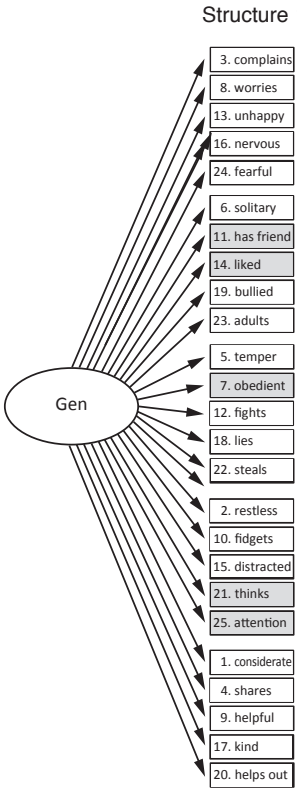


Figure 2. Hypothesised SDQ structures. Illustrations are simplified path diagrams, where rectangles represent SDQ items, ovals represent latent variables/factors, arrows from ovals to rectangles represent estimated factor loadings (i.e. those not fixed to zero) and curves between ovals represent estimated factor correlations (i.e. those not fixed to zero). Error variances of items are not shown, but they were estimated in all models. Covariances between errors were fixed to zero in all models.

b) Exploratory Structural Equation Modelling (ESEM)

Traditional CFA allows items to load only on specified (target) factors. These factor loadings are then estimated where the remaining non-target loadings are set to be precisely zero. This might be relatively restricting as low (but non-zero) item loadings might be present in the model. If they are, but forced to be zero, the issue translates into poor model fit. To address this, ESEM methodology has been developed.

Algorithms implemented in ESEM perform target rotation (Browne, 2001) on the pre-specified structure (in our case structure (1) depicted in Figure 2) but allow items to also load on other factors. This approach helps to investigate non-negligible loadings of items on other than target factors and thus the source of model misfit.

Sample

Descriptives for each sample are presented in Table 1. Around half of the children were of Pakistani origin, which reflects the birth profile of the city of Bradford, and half were female. A similar ethnic profile was seen across all three samples. There was a high rate of SDQ completion, and 70–80% of the parent completed versions were in English.

Table 1. Descriptives of each sample

	Age 3	Age 4	Age 5
N	1217	1711	2365
SDQ version used	Parent [2–4]	Parent [4–17]	Teacher [4–17]
Child age in years, mean (SD) [range]	3.1 (0.07) [2.9 to 3.4]	4.6 (0.34) [4.0 to 5.2]	5.2 (0.3) [4.5 to 5.9]*
missing N	0	1	26
Language of SDQ, N (%)			
English	971 (79.8)	1181 (69.0)	2365 (100)
Punjabi/Mirpuri	7 (0.6)	27 (1.6)	0
Urdu	240 (19.7)	503 (29.4)	0
missing, N	0	0	0
Child is female, N (%)	631 (51.9)	856 (50.0)	1194 (51.0)
Ethnicity of mother, N (% non-missing)			
White British	460 (37.8)	502 (29.4)	689 (36.3)
Pakistani	596 (49.1)	1016 (59.5)	974 (51.3)
Other	159 (13.1)	189 (11.1)	234 (12.3)
missing, N	2	4	468
Ethnicity of child**			
White British	452 (37.2)	495 (29.0)	771 (32.6)
Pakistani	596 (49.0)	1016 (59.6)	1205 (51.0)
Other	168 (13.8)	195 (11.4)	386 (16.3)
missing, N	1	2	3
All 25 SDQ items complete, N (%)	1212 (99.6)	1708 (99.9)	2304 (97.4)
Teachers, N	–	–	186
cases missing the teacher	–	–	3

Samples overlap. *two 6 year old children were removed from the dataset; **where school data are missing, child ethnicity is backfilled by the mothers' and may under-represent the number of 'other' ethnicity due to unknown mixed race

Confirmatory factor analysis (CFA)

The configuration with the most support for good fit was structure (1) in Figure 2 comprising a bifactor model of the 2 broader scales and a methods factor, using 20 items. The bifactor model accounts for common variance in all items and three specific factors. Two specific factors account for specific common variance within internalising and externalising items respectively. The third specific factor (methods factor) accounts for the five positively worded items. Importantly, our results suggest this model remains the most promising one across our three samples (Table 2).

Table 2. Fit indices of factor models

			FIML			WLSMV		
Configuration	**Age of sample**	**n**	**AIC**	**BIC**	**aBIC**	**CFI**	**RMSEA**	**WRMR**
(1) Bifactor: 2 Broader scales and Methods (20 items)	**3**	**1214**	**38818**	**39231**	**38974**	**0.967**	**0.036**	**1.280**
	4	**1709**	**49790**	**50231**	**49973**	**0.971**	**0.037**	**1.452**
	5	**2361**	**44027**	**44494**	**44236**	*	*	*
(2) Bifactor: 2 Broader scales (20 items)	3	1214	38863	39276	39019	0.952	0.043	1.455
	4	1709	49829	50270	50013	0.967	0.039	1.529
	5	2361	44076	44544	44286	0.989	0.055	2.286
(3) Bifactor: 5 Subscales and Methods (25 items)	3	1214	48007	48523	48202	*	*	*
	4	1709	60378	60927	60607	0.877	0.072	2.623
	5	2361	57105	57687	57366	0.988	0.060	2.54
(4) Bifactor: 5 Subscales (25 items)	3	1214	48078	48588	48271	0.879	0.067	2.153
	4	1709	60627	61171	60853	0.864	0.075	2.742
	5	2361	57104	57681	57363	0.987	0.060	2.566
(5) 5 Subscales and Methods (25 items)	3	1214	*	*	*	0.870	0.068	2.227
	4	1709	*	*	*	0.864	0.073	2.751
	5	2361	57938	58440	58164	*	*	*
(6) 4 difficulties Subscales (20 items)	3	1214	39007	39344	39134	0.918	0.054	1.798
	4	1709	50365	50725	50515	0.892	0.068	2.513
	5	2361	44160	44541	44331	0.983	0.066	2.809
(7) Second order (25 items)	3	1214	48224	48647	48384	0.857	0.071	2.320
	4	1709	60917	61368	61105	0.845	0.077	2.921
	5	2361	58133	58612	58348	*	*	*
(8) 2 Broader scales and Prosocial subscale (25 items)	3	1214	48357	48755	48507	0.839	0.074	2.446
	4	1709	61091	61515	61268	0.827	0.081	3.076
	5	2361	59694	60144	59896	*	*	*
(9) 2 Broader scales (20 items)	3	1214	39126	39437	39243	0.898	0.059	1.961
	4	1709	50537	50869	50676	0.872	0.073	2.714
	5	2361	45580	45932	45738	0.950	0.110	4.605

			FIML			WLSMV		
Configuration	**Age of sample**	**n**	**AIC**	**BIC**	**aBIC**	**CFI**	**RMSEA**	**WRMR**
(10) Unidimensional difficulties (20 items)	3	1214	39306	39612	39422	0.868	0.067	2.184
	4	1709	50928	51255	51064	0.835	0.082	3.049
	5	2361	47425	47771	47580	0.895	0.159	6.638
(11) Unidimensional (25 items)	3	1214	48956	49339	49100	0.779	0.087	2.818
	4	1709	61971	62380	62141	0.771	0.092	3.499
	5	2361	62050	62483	62244	0.923	0.141	6.043

Broader scales; Internalising and Externalising (10 items each); Subscales, emotional symptoms, peer problems, hyperactivity, conduct problems, prosocial (5 items each); bolded rows indicate the configuration with the best fit; FIML estimated using Full Information Maximum Likelihood; WLSMV estimated under Weighted Root Mean Square Residual
*Model has not converged

Factor loadings in Table 3 show how closely items are related to hypothesised factors in our best fitting model across all three samples. The loadings on factors tend to increase over time (age 3 to age 5) suggesting the SDQ becomes more structurally clear and thus more valid for older children. In addition, with the exception of the age 5 sample, items with different wording (methods items) show very low factor loadings on the general factor which can be interpreted as their differential wording substantially affecting their validity. Further, relatively low factor loadings on methods factors for the age 3 and age 4 samples show that most of the item variance is explained by the general factor. Finally, the age 5 sample shows high factor loadings on the externalising factor and low loadings on the internalising factor which suggest that the general factor interprets internalising problems.

Table 3. Standardised factor loadings of the best fitting model across three samples

	Subs-cale	General			Internalising			Externalising			Methods		
Item		**age 3**	**age 4**	**age 5**	**age 3**	**age 4**	**age 5**	**age 3**	**age 4**	**age 5**	**age 3**	**age 4**	**age 5**
Q03 complains	emo	0.545	0.623	0.163	0.038	–0.126	0.434	–	–	–	–	–	–
Q08 worries	emo	0.529	0.654	0.311	0.376	0.150	0.801	–	–	–	–	–	–
Q13 unhappy	emo	0.595	0.809	0.361	0.308	0.074	0.688	–	–	–	–	–	–
Q16 nervous	emo	0.248	0.263	0.349	0.270	0.287	0.726	–	–	–	–	–	–
Q24 fearful	emo	0.269	0.556	0.421	0.334	0.137	0.831	–	–	–	–	–	–
Q06 solitary	peer	0.177	0.248	0.791	0.476	0.538	0.167	–	–	–	–	–	–
Q11 has friend (M)	peer	0.066	–0.062	0.907	0.159	0.553	–0.061	–	–	–	0.463	0.446	0.237
Q14 liked (M)	peer	0.161	0.136	0.677	0.230	0.521	–0.024	–	–	–	0.451	0.452	0.417

Item	Subscale	General			Internalising			Externalising			Methods		
		age 3	age 4	age 5	age 3	age 4	age 5	age 3	age 4	age 5	age 3	age 4	age 5
Q19 bullied	peer	0.623	0.565	0.530	0.085	0.101	0.326	–	–	–	–	–	–
Q23 adults	peer	0.208	0.148	0.450	0.347	0.359	0.142	–	–	–	–	–	–
Q05 temper	con	0.555	0.548	0.377	–	–	–	0.299	0.345	0.598	–	–	–
Q07 obedient (M)	con	0.262	0.328	0.335	–	–	–	0.314	0.412	0.683	0.428	0.456	0.368
Q12 fights	con	0.546	0.492	0.251	–	–	–	0.153	0.269	0.650	–	–	–
Q18 lies	con	0.547	0.449	0.235	–	–	–	0.294	0.165	0.664	–	–	–
Q22 steals	con	0.652	0.284	0.263	–	–	–	0.124	0.066	0.499	–	–	–
Q02 restless	hyp	0.326	0.359	0.401	–	–	–	0.668	0.650	0.869	–	–	–
Q10 fidgets	hyp	0.314	0.429	0.417	–	–	–	0.594	0.667	0.860	–	–	–
Q15 distracted	hyp	0.348	0.400	0.448	–	–	–	0.462	0.471	0.787	–	–	–
Q21 thinks (M)	hyp	0.197	–0.049	0.449	–	–	–	0.231	0.255	0.618	0.447	0.518	0.366
Q25 attention (M)	hyp	0.191	0.063	0.466	–	–	–	0.249	0.366	0.651	0.446	0.498	0.340

M, positively phrased problem questions (methods); emo, emotional; con, conduct problems; hyp, hyperactive

Exploratory Structural Equation Modelling (ESEM)

Further investigation of the model and its misfit was done via ESEM. As current development of ESEM methodology does not allow for cross-loadings on specific factors, the method factors were removed. Results are presented in Table 4.

Table 4. Standardized bifactor ESEM factor loadings the best fitting model (without methods factor) across three samples

Item	Subscale	General			Internalising			Externalising		
		age 3	age 4	age 5	age 3	age 4	age 5	age 3	age 4	age 5
Q03 complains	emo	**0.538**	**0.654**	**–0.007**	**–0.082**	**–0.070**	**0.468**	0.101	–0.020	0.120
Q08 worries	emo	**0.670**	**0.643**	**0.036**	**0.122**	**0.227**	**0.854**	–0.104	–0.039	–0.049
Q13 unhappy	emo	**0.684**	**0.771**	**0.091**	**0.110**	**0.159**	**0.773**	–0.027	0.086	0.214
Q16 nervous	emo	**0.369**	**0.256**	**0.163**	**0.053**	**0.349**	**0.787**	0.024	–0.020	–0.057

	Subscale	General			Internalising			Externalising		
Q24 fearful	emo	**0.462**	**0.577**	**0.185**	**–0.005**	**0.255**	**0.914**	–0.093	–0.085	–0.071
Q06 solitary	peer	**0.335**	**0.231**	**0.701**	**0.133**	**0.451**	**0.403**	0.047	0.045	–0.201
Q11 has friend (M)	peer	**0.042**	**–0.143**	**0.826**	**0.553**	**0.646**	**0.247**	0.131	0.214	–0.073
Q14 liked (M)	peer	**0.120**	**0.001**	**0.692**	**0.813**	**0.695**	**0.204**	0.166	0.334	0.295
Q19 bullied	peer	**0.580**	**0.507**	**0.271**	**0.100**	**0.229**	**0.445**	0.045	0.091	0.189
Q23 adults	peer	**0.365**	**0.163**	**0.288**	**–0.028**	**0.269**	**0.316**	0.001	–0.003	–0.169
Q05 temper	con	**0.501**	**0.575**	**0.318**	–0.057	–0.013	0.173	**0.422**	**0.353**	**0.638**
Q07 obedient (M)	con	**0.118**	**0.263**	**0.460**	0.321	0.162	–0.047	**0.521**	**0.549**	**0.674**
Q12 fights	con	**0.429**	**0.455**	**0.175**	–0.021	0.063	0.006	**0.340**	**0.342**	**0.791**
Q18 lies	con	**0.419**	**0.445**	**0.138**	–0.016	0.063	0.038	**0.482**	**0.204**	**0.862**
Q22 steals	con	**0.535**	**0.153**	**0.137**	0.112	0.284	0.039	**0.302**	**0.194**	**0.643**
Q02 restless	hyp	**0.313**	**0.435**	**0.586**	–0.148	–0.168	–0.095	**0.614**	**0.569**	**0.733**
Q10 fidgets	hyp	**0.328**	**0.497**	**0.587**	–0.103	–0.079	–0.071	**0.554**	**0.585**	**0.734**
Q15 distracted	hyp	**0.361**	**0.404**	**0.624**	–0.071	0.089	0.011	**0.452**	**0.479**	**0.625**
Q21 thinks (M)	hyp	**0.095**	**–0.166**	**0.643**	0.278	0.239	0.015	**0.392**	**0.474**	**0.548**
Q25 attention (M)	hyp	**0.098**	**–0.058**	**0.672**	0.260	0.328	0.053	**0.402**	**0.610**	**0.527**

M, positively phrased problem questions (methods); emo, emotional; con, conduct problems; hyp, hyperactive; bolding indicates theorised factor

Non-target loadings on internalising and externalising factors are of particular interest. Figures in Table 4 suggest that positively-worded items load on non-target factors but this might be the consequence of the fact that the method factor has been removed from this analysis. Apart from that, Q22 (steals) and Q02 (restless) load positively and negatively respectively onto the internalising factor for the age 3 and age 4 samples and item Q05 (temper) loads onto the internalising factor for age 5, suggesting a fairly substantial internalising component of those items. Similarly, item Q13 (unhappy) has a fairly large positive loading on the externalising factor at age 5 whereas item Q23 (adults) and Q06 (solitary) load negatively on the externalising factor.

DISCUSSION

In this analysis, we aimed to assess construct validity and factorial structure of the SDQ across three parent or teacher rated samples of multi-ethnic children aged 3, 4 and 5. Of the 11 configurations tested, we found that a bifactor model comprising the 2 broader scales plus the methods factor showed the best fit across all three samples. Further investigation of this configuration under ESEM methodology indicated several items that continued to contribute to model misfit.

Most CFA analyses have been conducted in older age samples and, in contrast to our study, have reported good fit when including the prosocial scale (Stone et al., 2010). We tested five configurations that included the prosocial scale but found acceptable fit only in the two age 5 samples that converged on a solution, and poor fit across the age 3 and 4 samples. This could be due to differences in the structure by informant, because the age 5 sample were teacher rated (which tends to have higher reliability (Stone et al., 2010), and differences in item-level response by informant have been reported (Goodman, 2001; Mellor & Stokes, 2007). Or, there could be less relevance of this dimension to younger age children, or other sample-specific differences. In studies that have assessed both teacher and parent questionnaires, an Australian analysis of 914 children aged 7–17 years failed to find adequate fit for any configuration (Mellor & Stokes, 2007), and Hill and Hughes (2007) found only marginal baseline fit for a five-factor structure for either informant in their sample of US children (mean age six). Similarly, CFAs conducted using data from three-year old Spanish children found only marginal baseline fit for either the teacher or parent rated version in five-factor first order and second order configurations (Ezpeleta et al., 2013). Croft et al. (2015) found a less than adequate fit (CFI = 0.905) for a five-subscale structure but did not test solutions not involving the prosocial scale. It is difficult to unpick the reasons for the variation in fit between samples for solutions involving the prosocial scale as our study is unusual in that we tested different configurations with it included and excluded. We did this for pragmatic reasons, however, as the prosocial scale is not generally used when computing scores to assess the risk of any relevant psychopathology, we suggest this approach is repeated in other samples to confirm our findings.

We noted generally improved fit across all configurations and increased loadings on factors for older children, which may indicate that the structure of the SDQ becomes clearer as children mature and disordered behaviour becomes distinguishable from extreme but 'normal' behaviour. The children in BiB, however, are still relatively young, and the lack of acceptable fit for several hypothesised structures even at age 5 could be related to cultural variation in implicit item meanings and standards for behaviour in a multi-ethnic community. Due to small sample size we did not examine invariance by ethnicity, focussing instead on investigating features of the SDQ common across samples rather than trying to distinguish differences between them. The presence of cross-loading and low-loading items also precluded an examination of DIF. We employed a novel ESEM analysis that allowed us to explore items that contributed to misfit in the seemingly well-fitting model and found that in the 3 and 4 age samples, two externalising questions (Q02 (restless) and Q22 (steals/spiteful)) loaded onto the internalising factor, as did question Q05 (temper) in the age 5 sample. Three internalising items in the age

5 sample loaded onto the externalising factor (Q06 (solitary), Q13 (unhappy) and Q23 (adults)). Differences in analytic methods make comparison of misfitting items challenging between studies. At least superficially, some broad comparisons can be drawn, but these are tentative and sample sizes, population heterogeneity, and methods may have altered size and pattern of any or all loadings. Theunissen et al. (2013) using CFA in a parent-rated Dutch sample of 3–4 year olds found that Q03 (complains) and Q19 (bullied) low-loaded (<0.3) in a five-subscale configuration, while Croft, et al. (2015) reported a loading of 0.39 for Q21 (thinks) in their 3 year old sample. In their CFA study of Spanish 3 year olds Ezpeleta, et al. (2013) found that Q22 (steals/spiteful) low-loaded (<0.4) on the parent-rated version in both five-subscale and second-order models whereas Q19 (bullied) low-loaded (<0.4) on the teacher-rated version in both models. This may indicate some problems with items 22 (steals/spiteful, conduct subscale) and 19 (bullied, peer subscale), and potentially others, that need further investigation.

Other studies have also reported marginal improvements in fit by adding a methods factor to account for the effects of both positive wording of problem items and prosocial items (McCrory & Layte, 2012; van de Looij-Jansen et al., 2011). McCrory and Layte (2012) also examined, as we did, the effect of a five-item methods factor within the 20 difficulty items in parent-reported SDQs for nine year old Irish children, finding that the methods factor accounted for only 4% of the common variance. We noted that methods factor items tended to have lower loadings on the general factor for the age 3 and 4 samples than age 5, indicating a potentially stronger effect of item wording on the structure in samples of younger children. This needs to be confirmed in other studies. We found that, similar to others, specification of bifactor models resulted in improved fit (Caci et al., 2015; Stevanovic et al., 2014). This is to be expected because the SDQ was constructed as multi-subscale unidimensional measure which is reflected by the suggested methods of scoring (Goodman, 1997).

Our results have clear practical implications. First, they suggest that the validity of SDQ at a younger age (especially at age 3) is questionable and we would advise caution before using it to estimate psychopathological risk in children younger than 5 years of age. Further validation of the reliability and predictive validity in different samples and using multiple raters will improve our understanding of the performance of the SDQ in children less than 5 years old. The construct validity seems to improve with increasing age of the sample, although this aspect needs further exploration and cross validation. Second, the bifactor structure, found to be the best fitting in our sample, is suggestive of improvements in the way the SDQ is scored. Currently, each of the five subscales are scored by summing responses about five discrete behaviours, but our results generally do not support giving separate scores on each subscale. The bifactor model shows that both total scores and scores on internalising and externalising items make conceptual sense but may not be optimal as either does not take aspect of the other (i.e. the total score does not take into account the internalising and externalising factors, and part of the variance of internalising and externalising scores are due to the common (general) factor underlying them). In addition, neither scoring system does not take into account measurement error which is expressed (beside item error variance) as the method factor in our bifactor model. To fully acknowledge the underlying structure of SDQ we therefore suggest using factor scores of the proposed bifactor model instead of traditional SDQ scores.

One obvious caveat of this approach is that obtaining factor scores is not feasible for clinicians. For everyday practice, sum scores or mean scores are much more practical and provide instant information on level of risk of psychopathology. For everyday clinical use or for screening purposes where accuracy of scores is of less concern, we recommend using the total score and/or sum scores of the broader internalising and externalising scales. When larger samples or cohort data are available, we recommend re-estimation of our proposed bifactor model (or an alternative model if our results are not cross validated) and using factor scores in analysis.

Specifying a bifactor model has additional pragmatic advantages when applied to heterogeneous samples such as BiB where there is less support for the hypothesised structure. An individual score for the general factor can be generated which is controlled for the 'measurement error' resulting from low-loading and cross-loading items on the broader scales, and from positively worded items. These scores can then be used in comparative analyses, broadly interpreted as a total difficulties score e.g. (Prady et al., 2015). Alternatively, the internalising and externalising scores can be generated and interpreted as scores after the common variance to all items (externalising and internalising) has been accounted for by the general factor, and controlled for variation from the methods factor. Obviously, which scores are used (total, or broader scales) depends on the viewpoint of which components contain the 'nuisance' variation, and, for this reason, factor scores should be interpreted with caution. We suggest, however, that where data do not demonstrate the expected structure, whether due to multi-ethnicity or young age, and smaller sample sizes do not permit the exploration of the influence of these factors on the structure, that the extraction and use of factor scores represents a pragmatic solution that seeks to minimise measurement error. This is particularly pertinent as populations become increasingly heterogeneous, along with our desire to ever-more accurately assess behavioural problems in ever-younger children.

Limitations of our findings come mainly from the multi-ethnic nature of our sample. Clearly, validity of the SDQ may be different across ethnicities and therefore the scale might be prone to differential item functioning (DIF). We tried to investigate DIF in this sample but experienced notorious non-convergence of DIF models. This might be for two reasons: 1) relatively small sample sizes within each ethnicity compared to the complexity of the estimated model; and 2) structural ambiguity of the SDQ, especially in samples of younger children. In the future, larger studies should explore DIF properties of the SDQ in detail. We considered that DIF by ethnic group would be the largest source of variation, and so did not explore other potential sources such as informant sex, but the influence of such factors could also be usefully examined in larger samples. In the future, further development work in longitudinal samples containing independent diagnostic information could explore the utility of age-weighting SDQ data from younger children.

In conclusion, we found less support for the hypothesised structure and robustness of the SDQ in a multi-ethnic sample of 3, 4 and 5 year old children, but some evidence that the structure becomes more clear as children age from 3 to 5 years. We suggest that factor scores extracted from a bifactor model that account for measurement error could be useful if carefully applied in epidemiological and kinanthropological studies reporting risk of psychopathology in heterogeneous or younger age samples. We recommend that further work explore commonalities in measurement of child behaviour problems in multi-ethnic samples.

ACKNOWLEDGEMENTS

BiB has been possible only because of the enthusiasm and commitment of the children and parents who participated. The authors are grateful to all the participants, health professionals and researchers who made BiB happen.

FUNDING

This article presents independent research funded by the Medical Research Council, award reference MR/J013501/1, and the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH). JS was funded via Health e-Research Centre by the Medical Research Council, award reference MR/K006665/1 and partly by Charles University PRVOUK programme nr. P38. The views and opinions expressed are those of the authors, and not necessarily those of the Medical Research Council or the NIHR or the Department of Health. The funding bodies had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. All authors are independent of the funding bodies.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory.* Budapest: Akademiai Kiado.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bhugra, D. (2005). Cultural identities and cultural congruency: a new model for evaluating mental distress in immigrants. *Acta Psychiatr Scand*, *111*(2), 84–93.

Browne, M. W. (2001). An Overview of Analytic Rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research*, 2001/01/01, *36*(1), 111–150.

Caci, H., Morin, A. J., & Tran, A. (2015). Investigation of a bifactor model of the Strengths and Difficulties Questionnaire. *Eur Child Adolesc Psychiatry*, 2015/01/29, 1–11.

Croft, S., Stride, C., Maughan, B., & Rowe, R. (2015). Validity of the Strengths and Difficulties Questionnaire in Preschool-Aged Children. *Pediatrics*, May 1, 2015, *135*(5), e1210-e1219.

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457.

Ezpeleta, L., Granero, R., de la Osa, N., Penelo, E., et al. (2013). Psychometric properties of the Strengths and Difficulties Questionnaire(3–4) in 3-year-old preschoolers. *Compr Psychiatry*, Apr 2013, *54*(3), 282–291.

Fuchs, S., Klein, A. M., Otto, Y., & von Klitzing, K. (2013). Prevalence of emotional and behavioral symptoms and their impact on daily life activities in a community sample of 3 to 5-year-old children. *Child Psychiatry Hum Dev*, Aug 2013, *44*(4), 493–503.

Glazebrook, C., McPherson, A. C., Macdonald, I. A., Swift, J. A., et al. (2006). Asthma as a barrier to children's physical activity: implications for body mass index and mental health. *Pediatrics*, *118*(6), 2443–2449.

Goodman, A., Heiervang, E., Fleitlich-Bilyk, B., Alyahri, A., et al. (2012). Cross-national differences in questionnaires do not necessarily reflect comparable differences in disorder prevalence. *Soc Psychiatry Psychiatr Epidemiol*, Aug 2012, *47*(8), 1321–1331.

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. *J Abnorm Child Psychol*, Nov 2010a, *38*(8), 1179–1191.

Goodman, A., Patel, V., & Leon, D. A. (2010). Why do British Indian children have an apparent mental health advantage? *J Child Psychol Psychiatry*, Oct 2010b, *51*(10), 1171–1183.
Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *J Child Psychol Psychiatry*, Jul 1997, *38*(5), 581–586.
Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *J Am Acad Child Adolesc Psychiatry*, Nov 2001, *40*(11), 1337–1345.
Goodman, R., Ford, T., Simmons, H., Gatward, R., et al. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *Br J Psychiatry*, Dec 2000a, *177*, 534–539.
Goodman, R., Renfrew, D., & Mullick, M. (2000). Predicting type of psychiatric disorder from Strengths and Difficulties Questionnaire (SDQ) scores in child mental health clinics in London and Dhaka. *Eur Child Adolesc Psychiatry*, Jun 2000b, *9*(2), 129–134.
Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11 (Suppl 3)), 78.
Hamer, M., Stamatakis, E., & Mishra, G. (2009). Psychological distress, television viewing, and physical activity in children aged 4 to 12 years. *Pediatrics*, May 2009, *123*(5), 1263–1268.
Hill, C. R., & Hughes, J. N. (2007). An Examination of the Convergent and Discriminant Validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly*, *22*(3), 380–406.
Hu, L., & Bentler, M. P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.
Law, M., Petrenchik, T., King, G., & Hurley, P. (2007). Perceived Environmental Barriers to Recreational, Community, and School Participation for Children and Youth With Physical Disabilities. *Archives of Physical Medicine and Rehabilitation*, Dec 2007, *88*(12), 1636–1642.
Majnemer, A., Shevell, M., Law, M., Birnbaum, R., et al. (2008). Participation and enjoyment of leisure activities in school aged children with cerebral palsy. *Developmental Medicine & Child Neurology*, *50*(10), 751–758.
McCrory, C., & Layte, R. (2012). Testing competing models of the Strengths and Difficulties Questionnaire's (SDQ's) factor structure for the parent-informant instrument. *Personality and Individual Differences*, Jun 2012, *52*(8), 882–887.
Mellor, D., & Stokes, M. (2007). The Factor Structure of the Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment*, *23*(2), 102–112.
Mieloo, C. L., Bevaart, F., Donker, M. C., van Oort, F. V., et al. (2014). Validation of the SDQ in a multi-ethnic population of young children. *Eur J Public Health*, Feb 2014, *24*(1), 26–32.
Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In: K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models.* Newbury Park, CA: Sage, pp. 205–243.
Muthén, L., & Muthén, B. (2016). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén, 1998–2016a.
Muthén, L., & Muthén, B. (2016) *Mplus: Statistical analysis with latent variables*. [Version for 7.3]. Los Angeles, CA, 1998–2016b.
Nazroo, J. Y. (1998). Genetic, Cultural or Socio-economic Vulnerability? Explaining Ethnic Inequalities in Health. *Sociology of Health & Illness*, *20*(5), 710–730.
Nixon, G. M., Thompson, J. M., Han, D. Y., Becroft, D. M. et al. (2008). Short sleep duration in middle childhood: risk factors and consequences. *Sleep*, *31*(1), 71.
Page, A. S., Cooper, A. R., Griew, P., & Jago, R. (2010). Children's screen viewing is related to psychological difficulties irrespective of physical activity. *Pediatrics*, *126*(5), e1011–e1017.
Petermann, U., Petermann, F. and Schreyer, I. (2010). The German Strengths and Difficulties Questionnaire (SDQ): Validity of the teacher version for preschoolers. *European Journal of Psychological Assessment*, *26*(4), 256–262.
Prady, S. L., Pickett, K. E., Croudace, T., Mason, D., et al. (2015). Maternal psychological distress in primary care and association with child behavioural outcomes at age three. *Eur Child Adolesc Psychiatry*, Sep 2015, *24*, 1–13.
Raynor, P., & Born in Bradford Collaborative Group (2008). Born in Bradford, a cohort study of babies born in Bradford, and their parents: protocol for the recruitment phase. *BMC Public Health*, *8*, 327.

Sagatun, A., Søgaard, A. J., Bjertness, E., Selmer, R., et al. (2007). The association between weekly hours of physical activity and mental health: A three-year follow-up study of 15–16-year-old students in the city of Oslo, Norway. *BMC Public Health*, *7*(1), 1–9.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Steiger, J. H., & Lind, J. (1980). Statistically-based tests for the number of common factors. In: *Proceedings of the Annual Spring Meeting of the Psychometric Society*, Iowa City, IA 758.

Stevanovic, D., Urban, R., Atilola, O., Vostanis, P., et al. (2014). Does the Strengths and Difficulties Questionnaire – self report yield invariant measurements across different nations? Data from the International Child Mental Health Study Group. *Epidemiol Psychiatr Sci*, Apr 2014, 1–12.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., 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*, Sep 2010, *13*(3), 254–274.

Theunissen, M. H., Vogels, A. G., de Wolff, M. S., & Reijneveld, S. A. (2013). Characteristics of the strengths and difficulties questionnaire in preschool children. *Pediatrics*, Feb 2013, *131*(2), e446–454.

Ussher, M., Owen, C., Cook, D., & Whincup, P. (2007). The relationship between physical activity, sedentary behaviour and psychological wellbeing among adolescents. *Social Psychiatry and Psychiatric Epidemiology*, *42*(10), 851–856.

Van de Looij-Jansen, P. M., Goedhart, A. W., de Wilde, E. J., & Treffers, P. D. (2011). Confirmatory factor analysis and factorial invariance analysis of the adolescent self-report Strengths and Difficulties Questionnaire: how important are method effects and minor factors? *Br J Clin Psychol*, Jun 2011, *50*(2), 127–144.

Watson, K. D., Papageorgiou, A., Jones, G., Taylor, S., et al. (2003). Low back pain in schoolchildren: the role of mechanical and psychosocial factors. *Arch Dis Child*, *88*(1), 12–17.

Wiles, N., Jones, G., Haase, A., Lawlor, D., et al. (2008). Physical activity and emotional problems amongst adolescents. *Social Psychiatry and Psychiatric Epidemiology*, *43*(10), 765–772.

Wright, J., Small, N., Raynor, P., Tuffnell, D., et al. (2013). Cohort profile: The Born in Bradford multi-ethnic family cohort study. *Int J Epidemiol*, *42*(4), 978–991.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Doctoral dissertation, University of California, Los Angeles.

Jan Štochl
stochl@ftvs.cuni.cz