

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

Anna Goodman¹ Donna L. Lamping² and George B. Ploubidis¹

¹ Department of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

² Health Services Research Unit, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

Corresponding author: Anna Goodman, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK (anna.goodman@lshtm.ac.uk).

Note: this is a personal version, created by Anna Goodman, of the text of the accepted journal article. It reflects all changes made in the peer review process, but does not incorporate any minor modifications made at the proof stage. The complete citation for the final journal article is:

- Goodman, A; Lamping, D; Ploubidis, GB; (2010) When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *J Abnorm Child Psychol*, 38 (8). pp. 1179-91
- DOI: 10.1007/s10802-010-9434-x

Copyright © and Moral Rights for this paper are retained by the individual authors and/or other copyright owners

Abstract

The Strengths and Difficulties Questionnaire (SDQ) is a widely used child mental health questionnaire with five hypothesised subscales. There is theoretical and preliminary empirical support for combining the SDQ's hypothesised emotional and peer subscales into an 'internalizing' subscale and the hypothesised behavioral and hyperactivity subscales into an 'externalizing' subscale (alongside the fifth prosocial subscale). We examine this using parent, teacher and youth SDQ data from a representative sample of 5 to 16 year olds in Britain (N=18,222). Factor analyses generally supported second-order internalizing and externalizing factors, and the internalizing and externalizing subscales showed good convergent and discriminant validity across informants and with respect to clinical disorder. By contrast, discriminant validity was poorer between the emotional and peer subscales and between the behavioral, hyperactivity and prosocial subscales. This applied particularly to children with low scores on those subscales. We conclude that there are advantages to using the broader internalizing and externalizing SDQ subscales for analyses in low-risk samples, while retaining all five subscales when screening for disorder.

Key words: Strengths and Difficulties Questionnaire; Factor structure; Construct validity; Britain; internalizing problems; externalizing problems

Introduction

The Strengths and Difficulties Questionnaire (SDQ) is one of the most widely used brief questionnaires for assessing child mental health problems. In the decade since its development, it has been used in low-, middle- and high-income settings around the world (reviewed in Achenbach, et al., 2008; Woerner, Fleitlich-Bilyk, et al., 2004). The SDQ can be completed by parents and teachers of children aged 4 to 16 and by youth aged 11 to 16.

The SDQ consists of 25 items covering five subscales relating to emotional problems, peer problems, behavioral problems, hyperactivity and prosocial behavior (R. Goodman, 1997). The SDQ total difficulties score, which is the sum of the emotional, peer, behavioral and hyperactivity subscales, has been found to be a psychometrically sound measure of overall child mental health problems in studies from around the world (Achenbach, et al., 2008; A. Goodman & Goodman, 2009; R. Goodman, 1997, 1999; R. Goodman, Renfrew, & Mullick, 2000; R. Goodman & Scott, 1999; Klasen, et al., 2000; Mullick & Goodman, 2001). This includes evidence that the total difficulties score is correlated with existing questionnaire and interview measures, differentiates clinic and community samples, and is associated with increasing rates of clinician-rated diagnoses of child mental disorder across its full range.

Nevertheless, the internal structure of the SDQ is one area where there is ongoing controversy. The SDQ items and subscales were developed with reference to the main nosological categories recognised by contemporary classification systems of child mental disorders such as the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV; American Psychiatric Association, 1994). The five subscales were then refined through exploratory factor analyses (EFAs: R. Goodman, 1997), and have since been supported by EFAs in multiple samples from across Europe (Becker, et al., 2006; R. Goodman, 2001; Smedje, Broman, Hetta, & von Knorring, 1999; Woerner, Becker, & Rothenberger, 2004). Yet EFAs are an exploratory

technique, primarily useful in suggesting possible factor structures when these are not known. When a hypothesised factor structure exists, it is more appropriate to use a model-based framework such as confirmatory factor analysis (CFA: Brown, 2006).

Those CFAs which have been carried out provide at best mixed support for the SDQs five-factor structure. CFAs in Norway (youth SDQ) and Australia (parent, teacher and youth SDQ) found that models based on the hypothesised five factors did not show acceptable model fit for some or all indices considered (Mellor & Stokes, 2007; Ronning, Handegaard, Sourander, & Morch, 2004). Other CFAs in Belgium (parent and teacher SDQ) and Russia (youth SDQ) do report adequate global fit, but also note that loadings on several items were unacceptably low (<0.4) (Ruchkin, Kuposov, & Schwab-Stone, 2007; Van Leeuwen, Meerschaert, Bosmans, De Medts, & Braet, 2006).

This problematic evidence from CFAs suggests the possible value of considering alternative factor structures. One alternative which can be justified on theoretical grounds would combine the emotional and peer items into an ‘internalizing’ subscale and the behavioral and hyperactivity items into an ‘externalizing’ subscale. This approach receives some support from exploratory analyses; approximately internalizing/externalizing/prosocial factor structures have been reported in three-factor EFAs from the US (parent SDQ), Belgium (parent and teacher SDQ) and Finland (youth SDQ) (Dickey & Blumberg, 2004; Koskelainen, Sourander, & Vauras, 2001; Van Leeuwen, et al., 2006). A first-order model based on this three-factor solution showed adequate fit to the data in a CFA in the US sample, although the authors do not present CFAs of the five-factor solution for comparison (Dickey & Blumberg, 2004). By contrast, in the Belgium sample the three-factor solution did not achieve acceptable fit in a CFA and showed poorer fit the five-factor model (Van Leeuwen, et al., 2006).

These analyses therefore suggest that internalizing and externalizing factors may form part of the factor structure of the SDQ, but are not conclusive and do not investigate this issue in detail. There has also been no evaluation of other aspects of the construct validity of these theoretically plausible internalizing and externalizing subscales. Indeed, even for the five hypothesised SDQ subscales, almost all investigations of construct validity start and end with factor analyses such as those cited above. Far less use has been made of alternative approaches such as assessing convergent and discriminant validity – that is, the extent to which different subscale tap into *distinct* aspects of child mental health. Nevertheless such analyses have the potential to be highly informative in clarifying whether, or under what circumstances, these SDQ subscales are valid for use as screening devices for clinical disorder or as explanatory or outcome variables in epidemiological studies.

In this paper, we therefore compare different models whereby the hypothesised internalizing and externalizing subscales could form part of the factor structure of the SDQ. We then evaluate the convergent/discriminant validity of the internalizing/externalizing SDQ subscales, and compare their performance with the hypothesised five subscales. The two British surveys we use (the British Child and Adolescent Mental Health Surveys of 1999 and 2004) have not previously been used for these purposes, although other psychometric analyses (e.g. Cronbach’s alpha, principal component analyses) have been published for the earlier survey (R. Goodman, 2001).

Methods

Description of sample

The British Child and Adolescent Mental Health Surveys (B-CAMHS) were two nationally-representative surveys conducted in England, Scotland and Wales in 1999 and 2004. Children aged 5-15 years were sampled in B-CAMHS99 and 5-16 years in B-CAMHS04, using the British Child Benefit Register as a sampling frame; full details have been published elsewhere (Green, McGinnity, Meltzer, Ford, & Goodman, 2005; Meltzer, Gatward, Goodman, & Ford, 2000). Between the two B-CAMHS surveys, 26,544 children and adolescents were selected and their principal caregivers ('parents') were approached for face-to-face interview. Of these, 18,415 (69%) participated giving a sample which was 50.7% male with mean age 10.2 years. Parent SDQ data were available for 18,222 (99.0%) participants. With parental permission, teachers were also approached to participate (by postal questionnaire), as were the 11-16 year-olds themselves (by face-to-face interview). This resulted in SDQ data from 14,263 teachers (77.4% of participants) and 7,678 youth (91.9% of participants aged 11-16).

Both B-CAMHS surveys included a three-year follow-up. B-CAMHS99 followed-up all children with a disorder at baseline and a third of children with no disorder at baseline (Meltzer, Gatward, Corbin, Goodman, & Ford, 2003). B-CAMHS04 followed-up all children, regardless of disorder status at baseline (Parry-Langdon et al, 2008). In total, 11,222 children were selected for follow-up and 7,912 (70.5%) participated, giving a sample that was 51.7% male with mean age 13.2 years.

Description of measures

All participating parents, teachers and children were administered the Strengths and Difficulties Questionnaire (SDQ). As described above, this is a 25-item questionnaire with five hypothesised subscales: emotional problems, peer problems, behavioral problems, hyperactivity and prosocial behavior (R. Goodman, 1997, 2001). Each subscale comprises five questions with 3-point response scales ('Not true'=0, 'Somewhat true'=1, 'Certainly true'=2), with a subscale score range of 0-10. Ten of the 25 items are positively worded 'strengths'; these are reversed scored if they contribute to the emotional, peer, behavioral or hyperactivity subscales. In this paper, we also assess the construct validity of alternative ten-item 'internalizing' (emotional and peer items) and 'externalizing' subscales (behavioral and hyperactivity items) with ranges of 0-20. Throughout this paper, we excluded the small number of SDQs which were missing one or more subscale scores (<0.4% for parent, teacher and youth SDQs).

After completing the SDQ, all B-CAMHS participants completed the Development and Well-being Assessment (DAWBA). This is a detailed psychiatric interview administered by lay interviewers to parents and youth, with a briefer questionnaire for teachers (R. Goodman, Ford, Richards, Gatward, & Meltzer, 2000). Each section of the DAWBA uses skip-rules, one component of which is the relevant SDQ subscale; for example, the hyperactivity SDQ subscale for the hyperactivity disorder section. Each section begins with structured questions that cover the operationalised diagnostic criteria for DSM-IV (American Psychiatric Association, 1994). Structured questions are supplemented by open-ended questions which record verbatim a respondent's own description of problem areas. Clinicians review the closed and open responses

from all informants, identifying discrepancies within or between informants, and using the content, length and tone of the transcripts to interpret conflicting information (Meltzer, et al., 2000). On this basis, raters decide whether a particular child meets all the relevant DSM-IV criteria for an operationalised mental disorder. Raters can also assign 'Not Otherwise Specified' disorder, for example 'behavioral disorder, not otherwise specified' when children have substantial impairment from symptoms which do not quite meet operationalised criteria. In this paper, we group the mental disorders into emotional disorders (including anxiety and depressive disorders); behavioral disorders (including oppositional defiant and conduct disorder); attention-deficit/hyperactivity disorder (ADHD); and autistic spectrum disorders (ASD: including autism and Asperger syndrome). In British samples (including B-CAMHS), the DAWBA has been shown to have good inter-rater reliability (e.g. kappa 0.86 for inter-rater agreement for 'any mental disorder' in an epidemiological sample (Ford, Goodman, & Meltzer, 2003)). It also has good validity as judged against case-notes diagnoses, performs well in differentiating clinic/community samples, and shows strong associations with risk factors, service use and three-year prognosis (Ford, et al., 2003; R. Goodman, Ford, et al., 2000; Meltzer, et al., 2003).

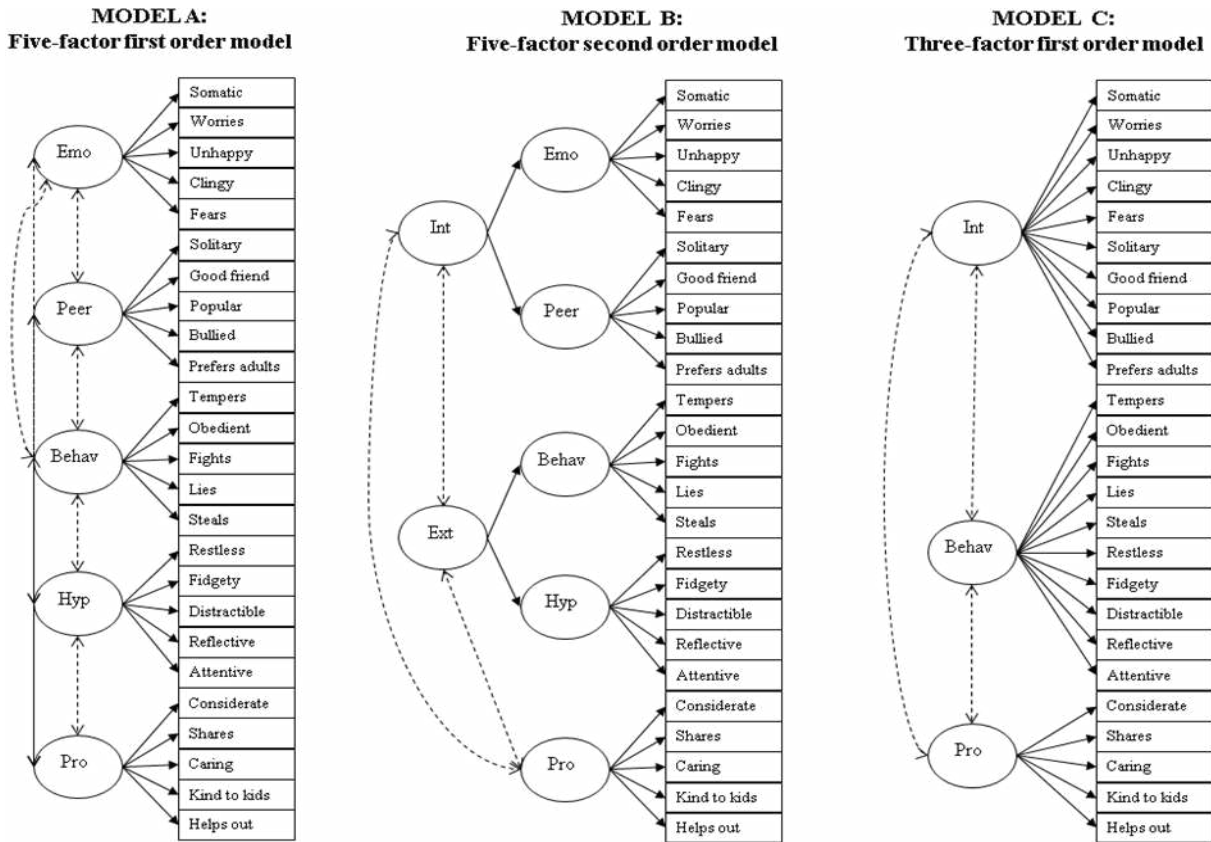
Statistical analyses

Factor structure of the SDQ

We used confirmatory factor analysis (CFA) to evaluate and compare the relative fit of a number of alternative factor structures for the parent, teacher and youth baseline SDQs. As shown in Figure 1, these were a first order model with the five hypothesised SDQ factors (Model A); a second order model with additional 'internalizing' and 'externalizing' factors (Model B); and a three-factor first order model in which internalizing and externalizing factors replaced the emotional, peer, behavioral and hyperactivity factors (Model C).

We performed the CFA in MPlus5, using a multivariate probit analysis for ordinal data (Muthen, 1983, 1984) and estimating model fit using the Weighted Least Squares, mean and variance adjusted (WLSMV) estimator. We follow common practice in reporting multiple indices of fit, namely the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI) and the Root Mean Square Error of Approximation (RMSEA) (Brown, 2006; Hu & Bentler, 1999). To consider a model as showing 'acceptable' fit, we required a CFI>0.90; TLI>0.90; and RMSEA<0.08; to consider a model as showing 'good' fit, we required a CFI>0.95; TLI>0.95; and RMSEA<0.06 (Brown, 2006). Where models showed acceptable fit on some indices but not on others, we allowed correlations between the unique variances of some individual items within the same factor, selecting these item pairs using MPlus' modification indices. Such minor model modifications can improve model fit by increasing the proportion of variance explained, but do not change the substantive conclusions regarding the adequacy of a hypothesised factor structure in describing a set of data (Bollen, 1989).

Figure 1: Models used in Confirmatory Factor Analyses of the parent, teacher and youth SDQ.



Construct validity of the SDQ subscales across informants

Multitrait-multimethod (MTMM) analyses are a method for assessing the construct validity of a set of measures (Campbell & Fiske, 1959; Nunnally & Bernstein, 1994). MTMM are based on a correlation matrix of multiple traits (e.g. the proposed SDQ subscales) measured by multiple methods (e.g. parent, teacher, youth). These can assess construct validity through comparisons across informants. For example, correlations between the parent and teacher behavioral subscales (a convergent correlation coefficient) would be expected to be higher than between the parent behavioral and teacher hyperactivity subscales (a discriminant correlation coefficient). If this aspect of construct validity could not be demonstrated, this would indicate that the behavioral and hyperactivity subscales are not tapping into the same, distinct constructs across informants.

We performed the MTMM analyses using subscales created by adding up the relevant items and *not* using the latent variables created through factor analyses. We did this because we believe that most users of the SDQ will prefer to use these simple, transparent scores, and that it is therefore their convergent and discriminant validity which it is most useful and most relevant to present. We assessed correlations between the (ordered) SDQ subscales using Spearman’s correlations, calculated in Stata 10.2 and basing each correlation coefficient upon all individuals with the relevant SDQ data. We also present the Cronbach alpha for each, as a measure of internal consistency.

Construct validity of the SDQ subscales relative to the DAWBA

MTMM analyses assess construct validity by comparing different informants. Comparing the SDQ and the DAWBA provides a further method of evaluating construct validity. The *a priori* prediction is that DAWBA diagnoses of emotional disorders should correlate most highly with the emotional SDQ subscale of the parent, teacher and youth SDQs; behavioral disorders with the behavioral subscale; ADHD with the hyperactivity subscale; and ASD with the peer and prosocial subscales. We performed a series of logistic regression analyses in Stata 10.2 on four outcomes: DAWBA diagnosis for any emotional disorder, any behavioral disorder, ADHD, or ASD. For the explanatory variables, we first used the five hypothesised SDQ subscales from the same informant. We then repeated these analyses using the three internalizing, externalizing and prosocial subscales. We reverse-scored the prosocial subscale for these analyses in order to facilitate comparisons of effect sizes across subscales.

Predicting baseline DAWBA diagnoses using baseline SDQ subscale scores is somewhat circular because the SDQ subscales form part of the skip rules for some DAWBA sections. High SDQ scores at baseline could therefore increase the probability of a DAWBA diagnosis at baseline simply by increasing the amount of mental health information collected. We therefore used DAWBA diagnoses at three-year follow-up, as these were administered and rated blind to SDQ score or DAWBA diagnosis at baseline. In doing so, we used weights to adjust for the fact that B-CAMHS99 did not seek to follow up all children but rather over-sampled children who had a disorder at baseline. We decided not to use the youth SDQ to predict ASD because only 10/71 children with a follow-up diagnosis of ASD completed youth SDQs at baseline, and these individuals may lack insight as informants.

Results

Internal factor structure of the SDQ

Table 1 presents the first-order model of the five hypothesised SDQ factors (Model A) for the parent, teacher and youth SDQs. Of the 75 standardised loadings (25 items times 3 informants), 37 were high (≥ 0.7) 36 were moderate (0.4-0.69) and only two ('good friend' and 'best with adults' on the youth SDQ) were unacceptably low (0.3-0.39). For all informants, Model A initially failed to demonstrate acceptable fit for at least one of the reported indices of global fit (CFI $<$ 0.90 for parents; RMSEA $>$ 0.08 for teachers; CFI and TLI $<$ 0.90 for youth). As reported in Table 2, just acceptable fit was usually achieved after allowing the unique variance to correlate between some items within the same factor, although the CFI in youth remained low (0.858). Taken together these results indicate that the hypothesised first order factor structure shows an 'acceptable' but not a 'good' fit to the parent, teacher and child SDQ data.

Table 1 also shows high correlations in all informants between the emotional and peer latent scores (0.66-0.71), and between the behavioral and hyperactivity subscales (0.71-0.81). This provides empirical support for our theory-driven intention to fit second-order internalizing and externalizing factors to capture these correlations, as shown in Model B. As is typical when comparing first-order and second-order models, there was relatively little difference between the fit of Model A and Model B – i.e. the second-order model showed a fit to the data which was 'acceptable' but generally not 'good'. This therefore supports the potential legitimacy of treating

internalizing and externalizing problems as broader factors subsuming the hypothesised subscales (although also highlights the fact that fitting this more complex model may not be necessary if one simply wishes to perform a CFA analyses to assess model fit). By contrast, replacing the emotional, peer, behavioral and hyperactivity factors with first order internalizing and externalizing factors (Model C) led to poorer model fit, indicating that this is not a legitimate simplification.

Table 1: Model fit and fully standardised item loadings from first order five-factor confirmatory factor analyses of the parent, teacher and youth SDQs (Model A)

		Parent	Teacher	Youth
N		18,222	14,263	7,678
MODEL FIT		CFI=0.857, TLI=0.934, RMSEA=0.059	CFI=0.905 TLI=0.963 RMSEA=0.085	CFI=0.837, TLI=0.885, RMSEA=0.063
STANDARDISED LOADINGS				
Factors	Items			
Emotional problems	Somatic	0.46	0.64	0.48
	Worries	0.68	0.78	0.66
	Unhappy	0.86	0.92	0.77
	Clingy	0.60	0.77	0.56
	Fears	0.70	0.84	0.67
Peer problems	Solitary	0.50	0.54	0.47
	Good friend*	-0.67	-0.80	-0.34
	Popular*	-0.82	-0.97	-0.58
	Bullied	0.67	0.58	0.73
	Best with adults	0.49	0.40	0.30
Behavioral problems	Tempers	0.67	0.77	0.66
	Obedient*	-0.71	-0.82	-0.59
	Fights	0.73	0.87	0.59
	Lies	0.72	0.86	0.70
	Steals	0.68	0.71	0.59
Hyperactivity problems	Restless	0.73	0.90	0.56
	Fidgety	0.78	0.91	0.65
	Distractible	0.80	0.90	0.74
	Reflective*	-0.69	-0.88	-0.59
	Persistent*	-0.75	-0.88	-0.65
Prosocial behavior	Considerate*	0.82	0.92	0.76
	Shares*	0.71	0.80	0.56
	Caring*	0.66	0.85	0.66
	Kind to kids*	0.68	0.80	0.66
	Helps out*	0.52	0.69	0.59
CORRELATION OF SUBSCALES		E with P: 0.71 E with B: 0.51 E with H: 0.40 E with Pr: -0.26 P with B: 0.58 P with H: 0.49 P with Pr: -0.47 B with H: 0.71 B with Pr: -0.70 H with Pr -0.50	E with P: 0.66 E with B: 0.34 E with H: 0.33 E with Pr: -0.24 P with B: 0.67 P with H: 0.54 P with Pr: -0.67 B with H: 0.81 B with Pr: -0.82 H with Pr -0.70	E with P: 0.69 E with B: 0.53 E with H: 0.48 E with Pr: -0.02 P with B: 0.47 P with H: 0.38 P with Pr: -0.45 B with H: 0.85 B with Pr: -0.54 H with Pr -0.49

Results from Model A, as defined in Figure 1. * indicates positively worded 'strengths' items. E=emotional latent score, P=peer latent score, B= behavioral latent score, H=hyperactivity latent score, Pr=prosocial latent score.

Table 2: Model fit in Confirmatory Factor Analyses of the parent, teacher and youth SDQs

		CFI	TLI	RMSEA
Parent (N=18,222)	Model A	0.857	0.934	0.059
	Model A, plus minor modifications†	0.901	0.954	0.049
	Model B, plus minor modifications†	0.900	0.953	0.049
	Model C, plus minor modifications†	0.871	0.938	0.057
Teacher (N=14,263)	Model A	0.905	0.963	0.085
	Model A, plus minor modifications†	0.919	0.970	0.077
	Model B, plus minor modifications†	0.921	0.969	0.078
	Model C, plus minor modifications†	0.877	0.948	0.101
Youth (N=7,678)	Model A	0.837	0.885	0.063
	Model A, plus minor modifications†	0.858	0.900	0.059
	Model B, plus minor modifications†	0.860	0.901	0.058
	Model C, plus minor modifications†	0.838	0.885	0.063

Models A, B and C defined in Figure 1. †Parent minor modifications: allowing correlation between the unique variance of (Clingy & Fears) (Solitary & Best with adults) (Restless & Fidgety) (Distractible & Persistent) (Reflective & Persistent). Teacher minor modifications: allowing correlation between the unique variance of (Worries & Fears) (Clingy & Fears) (Solitary & Best with adults) (Restless & Fidgety). Youth minor modifications: allowing correlation between the unique variance of (Restless & Fidgety)

Construct validity of the SDQ subscales across informants

Table 3 presents an MTMM analysis of the five SDQ subscales, created by summing the relevant five items from the parent, teacher and youth SDQs. The Cronbach alpha coefficients were almost all 0.65-0.85, indicating good internal reliability; the two exceptions were the peer problems subscales reported by parents ($\alpha=0.58$) and youth ($\alpha=0.44$). The cross-method correlations of the same traits are presented in bold; all were significantly different from zero ($p<0.001$) but were only low to moderate in magnitude (0.20-0.47). These convergent correlations were therefore similar in magnitude to the correlations between different subscales from the same informant.

In most cases the convergent correlations were significantly larger ($p<0.01$) than the other correlation coefficients in the same row or column (the discriminant correlations). There were, however, two important exceptions. First, in all three informant pairs, behavioral disorders did not show good discriminant validity relative to hyperactivity problems (relevant cells circled with solid line). For example, the correlation of parent behavioral and teacher behavioral scores was 0.31, no higher than the correlation between parent behavioral and teacher hyperactivity scores (0.31) and slightly lower than the correlation between parent hyperactivity and teacher behavioral scores (0.33). Second, the teacher prosocial subscale did not show discriminant validity relative to the behavioral and hyperactivity subscales reported by either parents or youth (relevant cells circled with dashed line).

Table 3: MTMM analyses for the five hypothesised SDQ subscales

		Parent					Teacher					Youth				
		Emo	Peer	Behav	Hyp	Pro	Emo	Peer	Behav	Hyp	Pro	Emo	Peer	Behav	Hyp	Pro
Parent	Emo	$\alpha=0.68$														
	Peer	0.37	$\alpha=0.58$													
	Behav	0.29	0.28	$\alpha=0.65$												
	Hyp	0.26	0.26	0.49	$\alpha=0.77$											
	Pro	-0.12	-0.17	-0.40	-0.32	$\alpha=0.66$										
Teacher	Emo	0.24	0.20	0.12	0.14	-0.05	$\alpha=0.78$									
	Peer	0.14	0.28	0.17	0.19	-0.13	0.41	$\alpha=0.69$								
	Behav	0.03	0.15	0.31	0.33	-0.18	0.18	0.36	$\alpha=0.75$							
	Hyp	0.07	0.17	0.31	0.47	-0.19	0.25	0.33	0.60	$\alpha=0.88$						
	Pro	-0.05	-0.15	-0.25	-0.28	0.25	-0.16	-0.40	-0.56	-0.57	$\alpha=0.84$					
Youth	Emo	0.36	0.20	0.14	0.13	-0.02	0.20	0.14	0.03	0.04	0.00	$\alpha=0.65$				
	Peer	0.19	0.34	0.12	0.13	-0.05	0.16	0.22	0.08	0.08	-0.08	0.32	$\alpha=0.44$			
	Behav	0.19	0.15	0.42	0.37	-0.25	0.11	0.14	0.29	0.30	-0.24	0.33	0.21	$\alpha=0.60$		
	Hyp	0.15	0.09	0.27	0.40	-0.17	0.13	0.09	0.25	0.33	-0.22	0.32	0.17	0.52	$\alpha=0.69$	
	Pro	-0.03	-0.08	-0.16	-0.17	0.30	-0.01	-0.09	-0.14	-0.16	0.24	-0.03	-0.15	-0.32	-0.30	$\alpha=0.66$

Emo=emotional SDQ subscale, peer=peer problems, behav=behavioral, hyp=hyperactivity, pro=prosocial. N=18,222 parents; N=14,263 teachers and N=7,678 youth. N=14139 for the parent-teacher comparison, N=7561 for the parent-youth comparison and N=5755 for the teacher-youth comparison. Values in cells are Spearman's correlation coefficients, except values in the diagonals which are Cronbach's alphas. Cross-method correlations of same traits are presented in bold. Cells circled with solid lines indicate problematic discriminant validity for the behavioral subscale relative to the hyperactivity subscale. Cells circled with dashed lines indicate problematic discriminant validity for the prosocial subscale relative to the behavioral and hyperactivity subscales

Table 4: MTMM analyses for the internalizing, externalizing and prosocial SDQ subscales

		Parent			Teacher			Youth		
		Int	Ext	Pro	Int	Ext	Pro	Int	Ext	Pro
Parent	Int	$\alpha=$ 0.73								
	Ext	0.37	$\alpha=$ 0.78							
	Pro	-0.18	-0.40	$\alpha=$ 0.66						
Teacher	Int	0.30	0.22	-0.11	$\alpha=$ 0.80					
	Ext	0.14	0.48	-0.21	0.36	$\alpha=$ 0.88				
	Pro	-0.11	-0.31	0.25	-0.32	-0.62	$\alpha=$ 0.84			
Youth	Int	0.40	0.18	-0.04	0.25	0.08	-0.04	$\alpha=$ 0.66		
	Ext	0.20	0.48	-0.23	0.15	0.37	-0.26	0.37	$\alpha=$ 0.76	
	Pro	-0.07	-0.19	0.30	-0.06	-0.16	0.24	-0.09	-0.35	$\alpha=$ 0.66

Int=internalizing, ext=externalizing, pro=prosocial SDQ subscales. N=18,222 parents; N=14,263 teachers and N=7,678 youth. N=14139 for the parent-teacher comparison, N=7561 for the parent-youth comparison and N=5755 for the teacher-youth comparison. Values in cells are Spearman's correlation coefficients, except values in the diagonals which are Cronbach's alphas. Cross-method correlations of same traits are presented in bold. Cells circled with dashed lines indicate problematic discriminant validity for the prosocial subscale relative to the externalizing subscales

The behavioral, hyperactivity and prosocial subscales therefore showed poor discriminant validity. Likewise the convergent correlations for the emotional and peer subscales were often not much larger than the discriminant correlations (although owing to the large sample size, all the differences were nonetheless significant at $p < 0.01$). These findings therefore do not support the claim that these five subscales tap into the same, *distinct* aspects of child mental health problems across all informants. By contrast, for the internalizing-externalizing contrast convergent and discriminant validity was much more satisfactory (see Table 4). However the prosocial subscale, particularly the teacher prosocial subscale, continued to show poor discriminant validity relative to the externalizing scale.

Construct validity of the SDQ subscales relative to the DAWBA

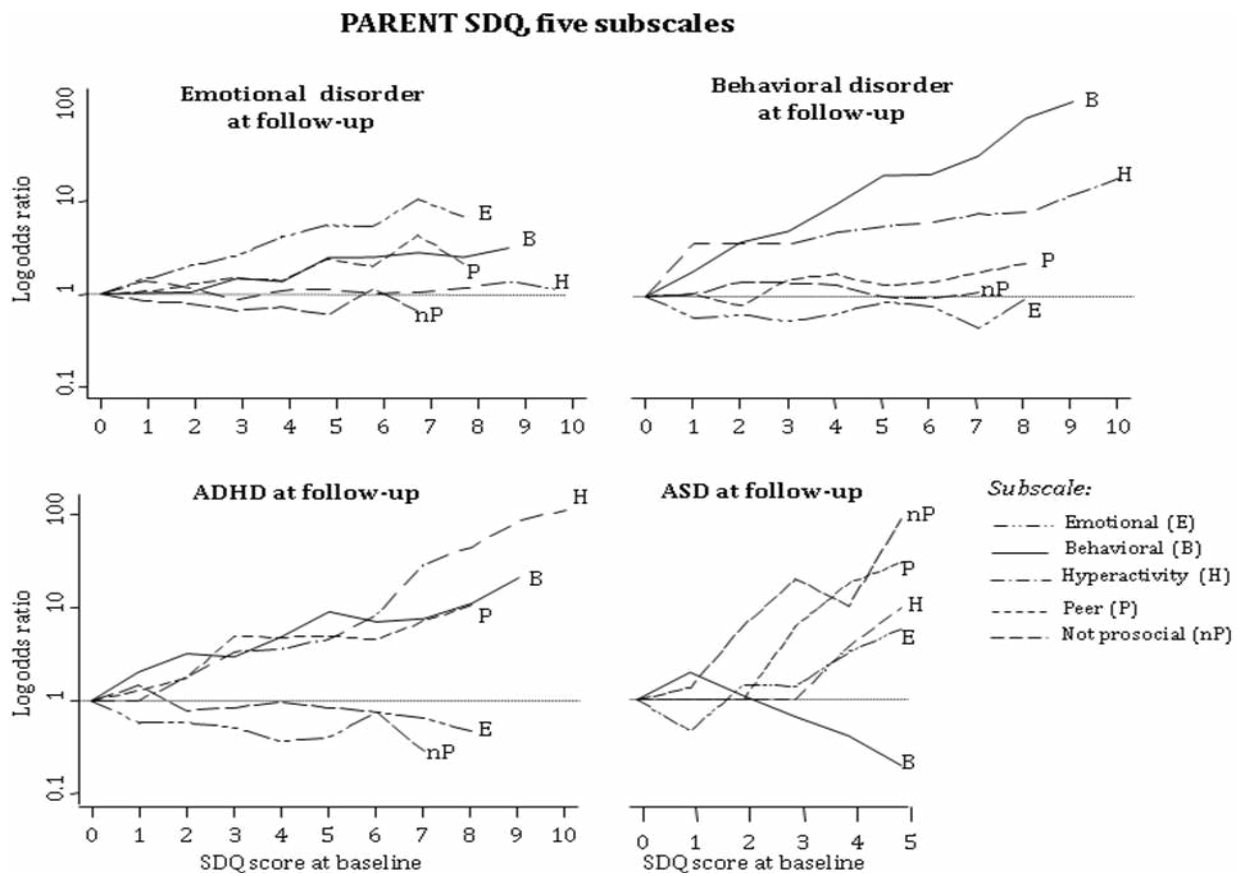
Both the baseline and the three-year follow-up prevalences of emotional, behavioral, ADHD and ASD generally showed monotonic increases across the full range of the corresponding parent, teacher and youth SDQ subscales at baseline (results available from www.sdqinfo.com/point_by_point.pdf). Among the five hypothesised SDQ subscales,

Table 5 shows which subscales had the largest effect upon the odds of receiving a DAWBA diagnoses at three-year follow-up (note that the prosocial subscale is reverse scored). For the parent and teacher SDQ, the expected subscale(s) always had the largest point estimates of effect size. These point estimates were also usually substantially and significantly larger than the next-largest estimates, except for the teacher emotional subscale (predicting to emotional disorder) and sometimes in the comparatively under-powered analyses predicting to ASD. For the youth SDQ evidence of discriminant validity was less convincing: the emotional subscale was no more strongly associated with emotional disorder than the peer subscale, and the hyperactivity subscale no more strongly associated with ADHD than the behavioral subscale.

The five-factor structure therefore generally showed convergent and discriminant validity relative to DAWBA diagnoses for parent and teacher SDQ but not always for the youth SDQ. Moreover, even for the parent and teacher SDQs, there was some suggestion that the behavioral and hyperactivity subscales only showed discriminant validity at higher scores. This is illustrated for the parent SDQ in Figure 2, which shows that below 7 SDQ points the behavioral and hyperactivity subscales were equally predictive of ADHD at follow-up. There is the suggestion of a similar effect below 3 SDQ points when predicting behavioral disorder at follow-up.

By contrast, the three-factor structure showed clear convergent and discriminant validity for all three informants (Table 6) and this was true even at the lowest SDQ scores. Graphs illustrating this can be found at www.sdqinfo.com/point_by_point.pdf, as can equivalent graphs to Figure 2 for the teacher and youth SDQs.

Figure 2: Independent association of the five parent SDQ subscales at baseline with DAWBA diagnoses at follow-up



Analyses come from models identical to those described in Table 5, except that the SDQ subscales were entered as categorical terms by SDQ point rather than as linear scales. Subscale scores were also grouped once the number of children per point fell to 20 or fewer, to avoid estimates based on very small numbers. As in Table 5, the prosocial score is reverse-scored to facilitate comparisons of effect sizes.

Table 5: Independent association of the five SDQ subscales at baseline with DAWBA diagnosis at follow-up (OR and 95% CI)

		Emotional DAWBA diagnosis	Behavioral DAWBA diagnosis	ADHD DAWBA diagnosis	ASD DAWBA diagnosis
Parents (N=7901)	Emotion (E)	1.32 (1.25, 1.39)***	0.99 (0.93, 1.05)	0.94 (0.85, 1.03)	1.25 (1.10, 1.43)**
	Peer (P)	1.14 (1.06, 1.22)***	1.09 (1.02, 1.17)*	1.29 (1.17, 1.43)***	1.58 (1.39, 1.80)***
	Behavioral (B)	1.16 (1.09, 1.25)***	1.65 (1.54, 1.76)***	1.33 (1.21, 1.46)***	0.65 (0.55, 0.76)***
	Hyperactivity (H)	1.00 (0.95, 1.05)	1.22 (1.16, 1.28)***	1.78 (1.60, 1.99)***	1.42 (1.25, 1.61)***
	Not Prosocial (nP)	0.93 (0.87, 1.00)*	1.02 (0.96, 1.09)	0.88 (0.79, 0.97)**	1.84 (1.61, 2.09)***
	<i>Largest subscale predictors</i>	<u>E</u> <u>B</u> <u>P</u> <u>H</u> <u>nP</u>	<u>B</u> <u>H</u> <u>P</u> <u>E</u> <u>nP</u>	<u>H</u> <u>B</u> <u>P</u> <u>E</u> <u>nP</u>	<u>nP</u> <u>P</u> <u>H</u> <u>E</u> <u>B</u>
Teachers (N=6247)	Emotion (E)	1.16 (1.09, 1.23)***	0.98 (0.92, 1.04)	0.91 (0.82, 1.01)	1.15 (1.02, 1.31)*
	Peer (P)	1.10 (1.02, 1.18)*	1.10 (1.03, 1.18)**	1.24 (1.12, 1.38)***	1.38 (1.21, 1.57)***
	Behavioral (B)	1.12 (1.02, 1.22)*	1.32 (1.23, 1.41)***	1.11 (1.01, 1.22)*	0.81 (0.66, 1.00)
	Hyperactivity (H)	1.01 (0.95, 1.07)	1.18 (1.12, 1.25)***	1.50 (1.38, 1.63)***	1.21 (1.05, 1.39)**
	Not Prosocial (nP)	1.01 (0.94, 1.07)	1.03 (0.96, 1.10)	1.00 (0.90, 1.10)	1.44 (1.21, 1.70)***
	<i>Largest subscale predictors</i>	<u>E</u> <u>B</u> <u>P</u> <u>H</u> <u>nP</u>	<u>B</u> <u>H</u> <u>P</u> <u>nP</u> <u>E</u>	<u>H</u> <u>P</u> <u>H</u> <u>nP</u> <u>E</u>	<u>nP</u> <u>P</u> <u>H</u> <u>E</u> <u>B</u>
Youth (N=3408)	Emotion (E)	1.24 (1.14, 1.34)***	0.97 (0.88, 1.06)	0.89 (0.72, 1.08)	–
	Peer (P)	1.25 (1.13, 1.37)***	1.06 (0.94, 1.19)	1.26 (0.98, 1.63)	–
	Behavioral (B)	1.04 (0.94, 1.16)	1.61 (1.44, 1.79)***	1.49 (1.18, 1.88)***	–
	Hyperactivity (H)	1.06 (0.98, 1.16)	1.13 (1.04, 1.23)**	1.30 (1.06, 1.58)*	–
	Not Prosocial (nP)	0.91 (0.83, 1.00)	0.99 (0.89, 1.10)	0.99 (0.83, 1.18)	–
	<i>Largest subscale predictors</i>	<u>P</u> <u>E</u> <u>B</u> <u>H</u> <u>nP</u>	<u>B</u> <u>H</u> <u>P</u> <u>nP</u> <u>E</u>	<u>B</u> <u>H</u> <u>P</u> <u>nP</u> <u>E</u>	

*p<0.05, **p<0.01, ***p<0.001. Odds ratios presented for probability of DAWBA diagnosis per one-point increase in the SDQ subscale in question. Below the odds ratios, the five subscales are presented in order of magnitude; subscales sharing an underline were not significantly different at p<0.05. Note that the prosocial score is reverse-scored to facilitate comparisons of effect sizes. ASD was not used as an outcome for the youth SDQ.

Table 6: Independent association of the three SDQ subscales at baseline with DAWBA diagnosis at follow-up (OR and 95% CI)

		Emotional DAWBA diagnosis	Behavioral DAWBA diagnosis	ADHD DAWBA diagnosis	ASD DAWBA diagnosis
Parents (N=7901)	Internalizing (In)	1.24 (1.20, 1.27)***	1.04 (1.00, 1.07)*	1.08 (1.02, 1.14)**	1.42 (1.34, 1.50)***
	Externalizing (Ex)	1.06 (1.02, 1.09)**	1.38 (1.33, 1.42)***	1.54 (1.45, 1.63)***	1.00 (0.94, 1.06)
	Not Prosocial (nP)	0.94 (0.87, 1.00)	1.07 (1.01, 1.13)*	0.88 (0.80, 0.97)**	1.74 (1.55, 1.95)***
	<i>Largest subscale predictors</i>	<u>In</u> <u>Ex</u> <u>nP</u>	<u>Ex</u> <u>nP</u> <u>In</u>	<u>Ex</u> <u>In</u> <u>nP</u>	<u>nP</u> <u>In</u> <u>Ex</u>
Teachers (N=6247)	Internalizing (In)	1.13 (1.10, 1.17)***	1.03 (1.00, 1.07)*	1.04 (0.99, 1.10)	1.24 (1.17, 1.32)***
	Externalizing (Ex)	1.05 (1.01, 1.09)*	1.23 (1.20, 1.28)***	1.32 (1.25, 1.39)***	1.02 (0.95, 1.10)
	Not Prosocial (nP)	1.00 (0.94, 1.07)	1.05 (0.98, 1.12)	1.01 (0.91, 1.12)	1.45 (1.21, 1.73)***
	<i>Largest subscale predictors</i>	<u>In</u> <u>Ex</u> <u>nP</u>	<u>Ex</u> <u>nP</u> <u>In</u>	<u>Ex</u> <u>In</u> <u>nP</u>	<u>nP</u> <u>In</u> <u>Ex</u>
Youth (N=3408)	Internalizing (In)	1.24 (1.18, 1.30)***	1.02 (0.97, 1.08)	1.03 (0.90, 1.18)	–
	Externalizing (Ex)	1.05 (1.01, 1.10)*	1.31 (1.24, 1.38)***	1.36 (1.22, 1.53)***	–
	Not Prosocial (nP)	0.91 (0.83, 1.00)	1.02 (0.92, 1.13)	1.04 (0.87, 1.25)	–
	<i>Largest subscale predictors</i>	<u>In</u> <u>Ex</u> <u>nP</u>	<u>Ex</u> <u>nP</u> <u>In</u>	<u>Ex</u> <u>nP</u> <u>In</u>	

*p<0.05, **p<0.01, ***p<0.001. Odds ratios presented for probability of DAWBA diagnosis per one-point increase in the SDQ subscale in question. Below the odds ratios, the three subscales are presented in order of magnitude; subscales sharing an underline were not significantly different at p<0.05. Note that the prosocial score is reverse-scored to facilitate comparisons of effect sizes. ASD was not used as an outcome for the youth SDQ.

Discussion

We used data from 18,222 British children demonstrate the construct validity of an ‘internalizing’ subscale (emotional plus peer items) and an ‘externalizing’ subscale (behavioral plus hyperactivity items) in the Strengths and Difficulties Questionnaire (SDQ). Second-order internalizing and externalizing factors were generally supported by confirmatory factor analyses, although model fit was somewhat problematic for the youth SDQ. The internalizing/externalizing subscales also showed the clearest and most consistent evidence of convergent and discriminant validity across informants and with respect to clinical disorder. By contrast, cross-informant discriminant validity was poorer between the emotional and peer subscales and particularly poor between the behavioral, hyperactivity and prosocial subscales. This suggests that in low-risk, epidemiological samples these five subscales may not all tap into distinct aspects of child mental health. Avoiding these five subscales and instead using the broader internalizing and externalizing subscales may therefore be more appropriate when selecting explanatory and outcome variables for epidemiological studies. Yet all five subscales on the parent and teacher SDQs did show convergent and discriminant validity when predicting to clinical disorder. This was particularly true for children with high scores on these subscales. As such, retaining all five subscales appears likely to add additional value when screening for disorder or studying high-risk children.

Our confirmatory factor analyses (CFAs) represent the first systematic evaluation of whether the parent, teacher and youth SDQs contain internalizing and externalizing factors, and of how these relate to the hypothesised five subscales. Our analyses did not support replacing the emotional, peer, behavioral and hyperactivity subscales with internalizing and externalizing factors. Instead this simplification produced worse model fit in all informants, thereby replicating the one previous study (of the parent and teacher SDQ) which made this comparison (Van Leeuwen, et al., 2006). By contrast, models which added second-order internalizing and externalizing factors did achieve acceptable values for all fit indices in the parent and teacher SDQ and for two out of three indices in the child SDQ. This provides some empirical support for our theoretically-driven proposal to evaluate the convergent and discriminant validity of the ten-item internalizing and externalizing SDQ subscales. Nevertheless, it should be noted that in all CFA analyses some indices of fit were ‘just acceptable’ rather than ‘good’. Moreover, on the youth SDQ the CFI index never achieved acceptable values and two item loadings were unacceptably low. These findings therefore add to the CFA evidence that the SDQ does not have a very clean internal factor structure (Mellor & Stokes, 2007) but that the hypothesised five subscales may nonetheless provide a passable description (Ronning, et al., 2004; Ruchkin, et al., 2007; Van Leeuwen, et al., 2006).

Our paper also extends the CFA literature by using additional approaches to evaluate construct validity. To our knowledge, this is the first time that full multitrait-multimethod (MTMM) analyses have been presented for the parent, teacher and youth SDQs. The convergent validity coefficients of 0.20-0.47 are lower than would be ideal, although this is typical in this respect of questionnaire measures of child psychopathology. For example, these values compare favourably to the inter-informant agreements reported in a meta-analysis of other child mental health questionnaires: 0.27 for parents and teachers, 0.25 between parents and children, 0.20 between teachers and children (Achenbach, McConaughy, & Howell, 1987). More worrying is the poor discriminant validity between the behavioral and hyperactivity subscales. This indicates that when applied to general population samples, the ‘behavioral’ and ‘hyperactivity’ labels may be misleading as these subscales cannot be assumed to be tapping into distinct aspects of externalizing problems. The MTMM analyses raised similar concerns for the

emotional vs. peer problems subscales, which likewise showed only weak evidence of cross-informant discriminant validity. The teacher prosocial subscale also did not show discriminant validity relative to the behavioral and hyperactivity subscales, suggesting that teachers may have been subsuming all these symptoms into a single ‘disruptive’/‘helpful’ continuum.

These findings suggest that it would not be valid (for example) to use mean scores from the behavioral and hyperactivity SDQ subscales in order to compare the correlates of behavioral vs. hyperactivity problems. If the same covariates were found to predict both subscales, then this might simply reflect the two subscales measuring the same thing rather than a real similarity in the correlates of behavioral and hyperactivity problems. Although firm recommendations are not possible without further replication, our provisional conclusion is therefore that the broader internalizing and externalizing subscales may be more appropriate explanatory or outcome variables in epidemiological studies. The internalizing and externalizing subscales also have the advantage that their greater number of items would be expected to reduce measurement error. This consideration may be particularly important when some populations of interest are small in size (e.g. minority ethnic groups).

Yet despite their poor cross-informant discriminant validity in MTMM analyses, all five SDQ subscales showed good discriminant validity when predicting clinical disorders. This seemed to be particularly true at higher SDQ subscale scores. One possible explanation for this discrepancy is that the MTMM analyses reflect patterns of subscale association in the *full* B-CAMHS sample, which is mostly comprised of children without mental health problems. In this low-risk, general population sample there may not always be a clear-cut distinction between (for example) behavioral and hyperactivity symptoms or between externalizing symptoms and prosocial behavior. Working with many children, teachers may find it particularly hard to make such distinctions, which could explain why discriminant validity between the externalizing and prosocial symptoms was particularly poor on the teacher SDQ. By contrast, discriminating symptom clusters may be easier when focusing on children with more severe mental health problems. An analogy from clinical practice would be the greater ease of distinguishing depressive and anxiety disorders in mental health specialist clinics than in the general population (Goldberg & Huxley, 1992).

We therefore conclude that there may be no single best set of subscales to use in the SDQ; rather, the optimal choice may depend in part upon one’s study population and study aims. Specifically, although the five hypothesised SDQ subscales should be treated with caution in low-risk samples, they do seem to add value when studying children with mental disorder and/or with higher SDQ scores. Strikingly, this applied not only to the emotional, behavioural and hyperactivity subscales when predicting the common child mental disorders, but also applied to the prosocial and peer problems subscales when predicting autistic spectrum disorders. Thus all five subscales appeared to have the potential to play a distinct, useful role when predicting child mental disorders, and this included subscales such as parent-reported peer problems which showed poor construct validity and internal reliability in the MTMM analyses. These findings are consistent with the fact that algorithms based on the five separate subscales have shown good performance in predicting type of disorder in clinics (R. Goodman, Renfrew, et al., 2000) or in the skip-rules of the DAWBA (R. Goodman, Ford, et al., 2000). They also highlight the vital importance of using multiple approaches to examine construct validity, and thereby building up a more complete and more nuanced picture of a measure’s performance. The unusually rich mental health data of our sample allowed us to go beyond most other studies in this regard, and we consider this a central strength of this paper.

Yet despite this key strength, our analyses and conclusions also have important limitations. The most important is the provisional nature of our conclusions regarding the optimal choice of SDQ subscales; firm recommendations must await replication in other studies. Other studies may also wish to use additional analytic approaches, such as conducting MTMM analyses within a CFA framework in order to estimate the convergent and discriminant correlation between the hypothesised latent trait (Brown, 2006). Although arguably less transparent than using the simple summed scores (hence our decision to use the 'traditional' approach in this paper), this would have the advantage of reducing measurement error. Finally, future studies could usefully be extended by including evidence from a larger number of domains of child psychopathology. These may be important in revealing aspects of convergent or discriminant validity for the SDQ subscales which are not apparent here. For example, factor analyses in an Australian sample of 4 to 9 year olds provide some evidence that parent-reported callous and unemotional traits (from a psychopathy measure) load with the prosocial SDQ items but not the behavioral or hyperactive items (Dadds, Fraser, Frost, & Hawes, 2005). This was not apparent in B-CAMHS04, however, where the magnitude of the correlation between the prosocial subscale and callous and unemotional traits was intermediate between the behavioral and hyperactivity subscales (Moran, et al., 2009). This discrepancy between the Australian sample and B-CAMHS04 further highlights the need for replication of our findings across other large datasets with multiple informants and high-quality diagnoses.

Conclusion

To summarise, the SDQ has several attractive features including a brief format, comparable versions for parents, teachers and young people, and versions in over 60 languages (see www.sdqinfo.com). These analyses add to the evidence, however, that the hypothesised five subscales may not always tap distinct constructs. Our analyses further indicate that the optimal choice of subscales may depend on one's study population and study aims. Our findings indicate that studies examining the broad constructs of internalizing and externalizing problems would be justified in using the SDQ to do so. Moreover, particularly in low-risk samples, this may be the more conservative approach in order to ensure an accurate description of what is being assessed and in order to generate findings which are comparable across informants. By contrast, using the five separate subscales may only be justified when seeking to study high-risk children, including those with mental disorder and/or with higher scores on the SDQ subscales.

Conflict of interest

AG is a director of YouthInMind, which provides no-cost and low-cost software and web sites related to the SDQ and the DAWBA.

References

- Achenbach, T. M., Becker, A., Dopfner, M., Heiervang, E., Roessner, V., Steinhausen, H. C., et al. (2008). Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: research findings, applications, and future directions. *J Child Psychol Psychiatry*, *49*(3), 251-275.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychol Bull*, *101*(2), 213-232.
- American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV)*. Washington, DC: American Psychiatric Association.
- Becker, A., Steinhausen, H. C., Baldursson, G., Dalsgaard, S., Lorenzo, M. J., Ralston, S. J., et al. (2006). Psychopathological screening of children with ADHD: Strengths and Difficulties Questionnaire in a pan-European study. *European Child and Adolescent Psychiatry*, *15*(Suppl. 1), 56-62.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. Guilford: Guilford Press.
- Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.
- Dadds, M. R., Fraser, J., Frost, A., & Hawes, D. J. (2005). Disentangling the Underlying Dimensions of Psychopathy and Conduct Problems in Childhood: A Community Study. *Journal of Consulting and Clinical Psychology*, *73*(3), 400-410.
- Dickey, W., & Blumberg, S. (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*(9), 1159-1167.
- Ford, T., Goodman, R., & Meltzer, H. (2003). The British Child and Adolescent Mental Health Survey 1999: the prevalence of DSM-IV disorders. *J Am Acad Child Adolesc Psychiatry*, *42*(10), 1203-1211.
- Goldberg, D., & Huxley, P. (1992). *Common mental disorders: a bio-social model*. London: Tavistock.
- Goodman, A., & Goodman, R. (2009). Strengths and difficulties questionnaire as a dimensional measure of child mental health. *J Am Acad Child Adolesc Psychiatry*, *48*(4), 400-403.
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *J Child Psychol Psychiatry*, *38*(5), 581-586.
- Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *J Child Psychol Psychiatry*, *40*(5), 791-799.
- Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *J Am Acad Child Adolesc Psychiatry*, *40*(11), 1337-1345.
- Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. *J Child Psychol Psychiatry*, *41*(5), 645-655.
- 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*, *9*(2), 129-134.
- Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: is small beautiful? *J Abnorm Child Psychol*, *27*(1), 17-24.
- Green, H., McGinnity, A., Meltzer, H., Ford, T., & Goodman, R. (2005). *Mental health of children and young people in Great Britain, 2004*: Palgrave MacMillan.
- Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

- Klasen, H., Woerner, W., Wolke, D., Meyer, R., Overmeyer, S., Kaschnitz, W., et al. (2000). Comparing the German versions of the Strengths and Difficulties Questionnaire (SDQ-Deu) and the Child Behavior Checklist. *Eur Child Adolesc Psychiatry*, 9(4), 271-276.
- Koskelainen, M., Sourander, A., & Vauras, M. (2001). Self-reported strengths and difficulties in a community sample of Finnish adolescents. *Eur Child Adolesc Psychiatry*, 10(3), 180-185.
- Mellor, D., & Stokes, M. (2007). The Factor Structure of the Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment*, 23(2), 105-112.
- Meltzer, H., Gatward, R., Corbin, T., Goodman, R., & Ford, T. (2003). *Persistence, onset, risk factors and outcomes of childhood mental disorders*. London: The Stationery Office.
- Meltzer, H., Gatward, R., Goodman, R., & Ford, T. (2000). *Mental health of children and adolescents in Great Britain*. London: The Stationery Office.
- Moran, P., Flach, C., Rowe, R., Briskman, J., Ford, T., Maughan, B., et al. (2009). Predictive value of callous-unemotional traits in a large community sample. *J Am Acad Child Adolesc Psychiatry*, 48(11), 1079-1084.
- Mullick, M. S., & Goodman, R. (2001). Questionnaire screening for mental health problems in Bangladeshi children: a preliminary study. *Soc Psychiatry Psychiatr Epidemiol*, 36(2), 94-99.
- Muthen, B. (1983). Latent Variable Structural Equation Modeling with Categorical-Data. *Journal of Econometrics*, 22(1-2), 43-65.
- Muthen, B. (1984). A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators. *Psychometrika*, 49(1), 115-132.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.
- Parry-Langdon et al. (2008). *Three years on: Survey of the development and emotional well-being of children and young people*. Cardiff: Office for National Statistics.
- Ronning, J. A., Handegaard, B. H., Sourander, A., & Mørch, W. T. (2004). The Strengths and Difficulties Self-Report Questionnaire as a screening instrument in Norwegian community samples. *Eur Child Adolesc Psychiatry*, 13(2), 73-82.
- Ruchkin, V., Kopolosov, R., & Schwab-Stone, M. (2007). The Strength and Difficulties Questionnaire: Scale validation with Russian adolescents. *Journal of Clinical Psychology*, 63, 861-869.
- Smedje, H., Broman, J. E., Hetta, J., & von Knorring, A. L. (1999). Psychometric properties of a Swedish version of the "Strengths and Difficulties Questionnaire". *Eur Child Adolesc Psychiatry*, 8(2), 63-70.
- Van Leeuwen, K., Meerschaert, T., Bosmans, G., De Medts, L., & Braet, C. (2006). The Strengths and Difficulties Questionnaire in a Community Sample of Young Children in Flanders. *European Journal of Psychological Assessment*, 22(3), 189-197.
- Woerner, W., Becker, A., & Rothenberger, A. (2004). Normative data and scale properties of the German parent SDQ. *Eur Child Adolesc Psychiatry*, 13 Suppl 2, II/3-10.
- Woerner, W., Fleitlich-Bilyk, B., Martinussen, R., Fletcher, J., Cucchiaro, G., Dalgalarondo, P., et al. (2004). The Strengths and Difficulties Questionnaire overseas: evaluations and applications of the SDQ beyond Europe. *Eur Child Adolesc Psychiatry*, 13 Suppl 2, II/47-54.

Supplementary material

In the first part of this supplementary material (page 21), we demonstrate that individual subscales of the SDQ are truly dimensional measures of mental health problems. This extends our recent demonstration that this is true of the total difficulty scores of the parent, teacher and youth SDQ (Goodman, A. and R. Goodman, *Strengths and difficulties questionnaire as a dimensional measure of child mental health*. J Am Acad Child Adolesc Psychiatry, 2009. **48**(4): p. 400-3.) In the second part of this report (page 29), we illustrate graphically the results of logistic regression analyses which investigate which SDQ subscale(s) are most strongly associated with different types of DAWBA diagnoses at three-year follow-up.

In both these supplementary analyses, we present results using the following subscales derived from the SDQ:

- The SDQ's five hypothesised subscales, emotional problems, peer problems, behavioral problems, hyperactivity and prosocial behaviour. These have a range from 0-10.
- The SDQ's alternative 'internalizing' (emotional plus peer subscales) and 'externalizing' (behavioural plus hyperactivity) subscales. These have a range of 0-20
- The SDQ's autism screening score, calculated as the peer problems subscale score minus the prosocial subscale. This has a range of -10 to +10

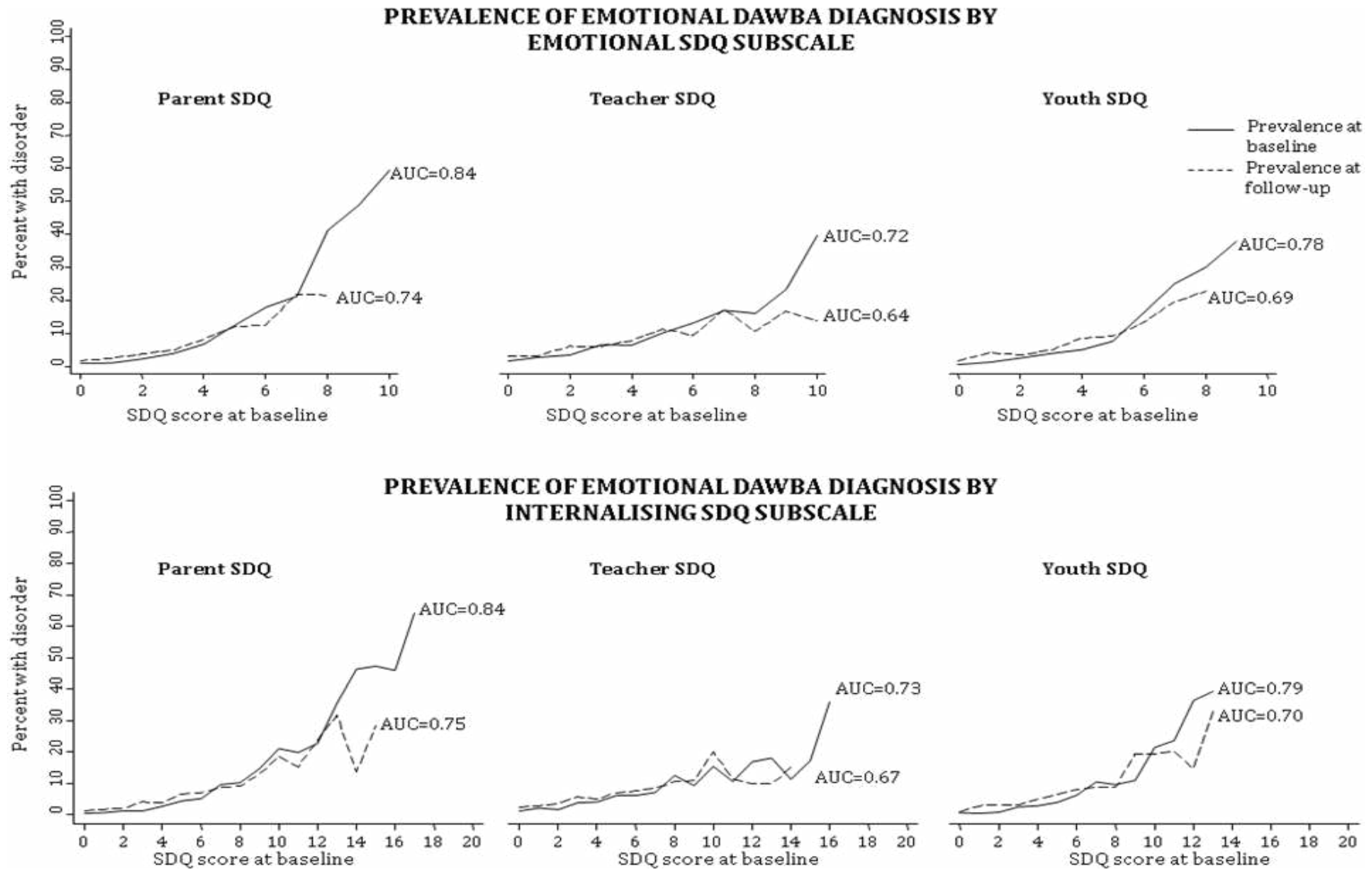
PART 1: Dimensionality of the subscales of the Strengths and Difficulties Questionnaire (SDQ)

Table 7: Prevalence of emotional disorder at baseline and follow-up by the parent, teacher and youth emotional and internalising SDQ subscales at baseline

		EMOTIONAL DISORDER					
		Parent SDQ		Teacher SDQ		Youth SDQ	
		Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI	Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI	Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI
		N	7901	14263	6247	7678	3408
Emotion-al subscale score	AUC	0.84	0.74	0.72	0.64	0.78	0.69
	0	0.5 (0.7, 1.0)	1.6 (1.2, 2.2)	1.6 (1.3, 1.9)	2.8 (2.2, 3.4)	0.7 (0.3, 1.4)	1.7 (0.8, 3.5)
	1	0.7 (0.9, 1.2)	2.4 (1.8, 3.2)	2.8 (2.2, 3.4)	3.0 (2.2, 4.3)	1.3 (0.9, 2.1)	3.7 (2.6, 5.3)
	2	1.7 (2.1, 2.7)	3.6 (2.7, 4.8)	3.4 (2.6, 4.4)	5.7 (4.2, 7.7)	2.6 (1.9, 3.5)	2.9 (1.8, 4.4)
	3	3.0 (3.7, 4.6)	4.8 (3.6, 6.5)	6.5 (5.3, 8.0)	4.9 (3.4, 7.0)	4.0 (3.0, 5.3)	4.6 (3.2, 6.6)
	4	5.4 (6.6, 8.2)	8.3 (6.2, 10.9)	6.4 (4.9, 8.4)	6.7 (4.3, 10.3)	5.1 (3.9, 6.8)	7.9 (5.6, 10.9)
	5	10.4 (12.4, 14.7)	11.9 (9.0, 15.6)	10.0 (7.6, 13.1)	9.5 (6.4, 13.7)	7.7 (5.9, 9.9)	8.6 (5.8, 12.6)
	6	14.8 (17.8, 21.2)	12.4 (8.9, 17.1)	13.2 (10.0, 17.2)	10.3 (5.5, 18.4)	16.3 (12.9, 20.2)	12.5 (8.2, 18.5)
	7	17.0 (21.1, 26.0)	21.8 (15.3, 30.0)	16.9 (12.2, 23.1)	13.8 (7.7, 23.5)	25.1 (19.2, 32.0)	17.0 (9.9, 27.6)
	8	33.6 (41.1, 49.1)	21.5 (15.9, 28.5)	16.0 (10.5, 23.8)	11.6 (5.3, 23.6)	30.0 (22.1, 39.2)	20.5 (12.2, 32.5)
	9	38.7 (48.6, 58.7)		23.4 (14.5, 35.3)	13.1 (5.0, 30.4)	37.8 (25.8, 51.4)	
10	44.9 (59.2, 72.2)		39.8 (25.7, 55.7)	11.0 (3.1, 32.0)			
Internalising subscale score	AUC	0.84	0.75	0.73	0.67	0.79	0.70
	0	0.5 (0.3, 0.8)	1.1 (0.7, 1.9)	1.2 (0.9, 1.6)	2.2 (1.6, 3.0)	0.9 (0.3, 2.4)	1.0 (0.2, 4.2)
	1	0.7 (0.5, 1.1)	1.7 (1.1, 2.7)	2.2 (1.7, 2.8)	2.8 (1.9, 4.1)	0.6 (0.2, 1.4)	2.4 (1.3, 4.7)
	2	1.4 (1.0, 1.8)	2.1 (1.5, 3.0)	1.7 (1.2, 2.4)	3.3 (2.3, 4.7)	1.0 (0.6, 1.7)	3.0 (1.9, 4.7)
	3	1.3 (0.9, 1.9)	4.1 (3.1, 5.5)	3.9 (3.0, 5.1)	5.2 (3.7, 7.2)	2.7 (1.9, 3.9)	2.7 (1.6, 4.5)
	4	2.8 (2.1, 3.6)	3.9 (2.6, 5.7)	4.1 (3.1, 5.5)	4.2 (2.6, 6.6)	3.0 (2.1, 4.3)	4.4 (2.9, 6.7)
	5	4.5 (3.5, 5.7)	6.6 (4.8, 9.2)	6.2 (4.7, 8.1)	6.6 (4.4, 9.7)	4.1 (2.9, 5.7)	6.3 (4.1, 9.5)
	6	5.2 (4.0, 6.7)	6.9 (5.0, 9.4)	6.1 (4.5, 8.3)	6.2 (3.9, 9.7)	6.4 (4.7, 8.6)	7.3 (4.7, 11.1)
	7	9.6 (7.7, 11.9)	8.7 (6.1, 12.2)	7.2 (5.2, 9.8)	7.2 (4.3, 11.8)	10.6 (8.2, 13.6)	7.8 (4.8, 12.3)
	8	10.2 (8.0, 13.0)	9.1 (6.1, 13.4)	12.5 (9.6, 16.2)	8.6 (5.4, 13.3)	9.9 (7.3, 13.1)	8.4 (4.8, 14.2)
	9	14.7 (11.7, 18.4)	12.9 (8.4, 19.4)	9.4 (6.5, 13.4)	8.9 (5.3, 14.7)	11.1 (7.5, 16.2)	17.2 (10.8, 26.1)
	10	21.1 (16.9, 26.1)	18.6 (12.7, 26.5)	15.5 (11.0, 21.4)	17.4 (10.6, 27.3)	21.6 (15.8, 28.8)	17.1 (9.9, 27.8)
	11	19.9 (14.6, 26.5)	15.3 (9.7, 23.5)	10.7 (6.4, 17.4)	10.0 (4.3, 21.8)	23.8 (16.4, 33.3)	21.6 (12.0, 35.7)
	12	22.8 (16.7, 30.3)	23.8 (14.1, 37.2)	16.9 (11.0, 25.1)	11.0 (4.1, 26.2)	36.6 (26.0, 48.6)	12.8 (4.2, 33.0)
	13	35.7 (26.2, 46.4)	31.7 (17.9, 49.8)	18.1 (11.3, 27.7)	9.6 (3.2, 25.4)	39.4 (29.4, 50.5)	30.6 (16.6, 49.4)
	14	46.5 (35.9, 57.4)	13.8 (5.6, 30.3)	11.4 (5.5, 22.2)	15.0 (8.1, 26.0)		
	15	47.4 (33.9, 61.3)	28.6 (18.0, 42.1)	17.2 (7.3, 35.6)			
	16	46.1 (30.5, 62.5)		36.0 (25.1, 48.6)			
	17	64.3 (47.4, 78.3)					
	18						
19							
20							

AUC=area under the curve, determined by receiver operating characteristic analyses. High scores were grouped once the number of children per SDQ point fell to 20 or fewer to prevent very small numbers leading to instability in the prevalence estimates. The last prevalence estimate presented for each informant and at each time point therefore corresponds to the prevalence for that number of SDQ points or above.

Figure 3: Prevalence of emotional disorder at baseline and follow-up by the parent, teacher and youth emotional and internalising SDQ subscales at baseline



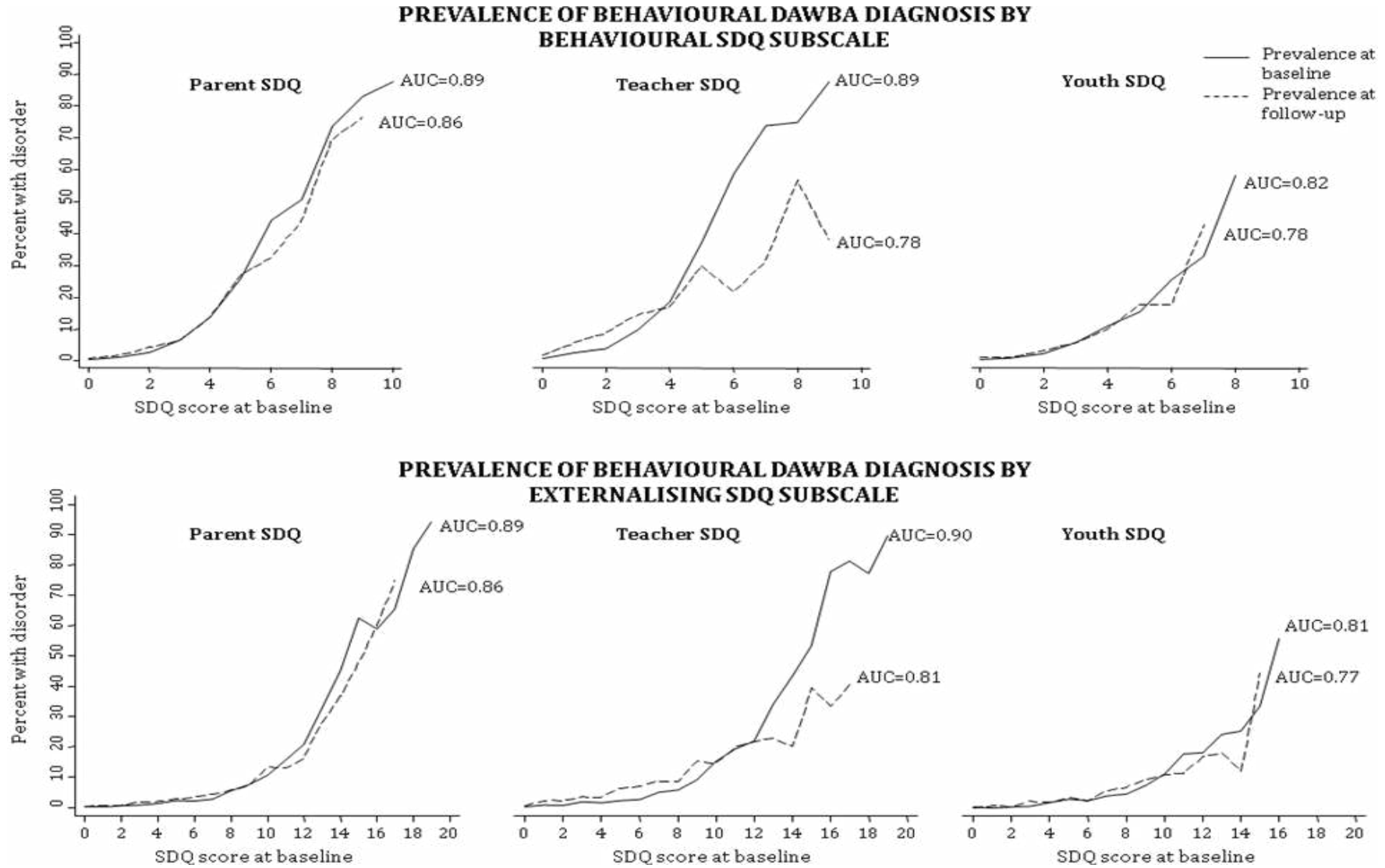
AUC=area under the curve

Table 8: Prevalence of behavioural disorder at baseline and follow-up by the parent, teacher and youth behavioural and externalising SDQ subscales at baseline

		BEHAVIOURAL DISORDER					
		Parent SDQ		Teacher SDQ		Youth SDQ	
		Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI	Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI	Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI
	N	18222	7901	14263	6247	7678	3408
Behav- ioural subscale score	AUC	0.89	0.86	0.89	0.78	0.82	0.78
	0	0.3 (0.2, 0.5)	0.7 (0.5, 1.1)	0.8 (0.6, 1.0)	1.6 (1.3, 2.1)	0.4 (0.2, 1.0)	1.1 (0.5, 2.3)
	1	1.1 (0.9, 1.5)	1.7 (1.2, 2.4)	2.6 (1.9, 3.4)	5.0 (3.7, 6.6)	1.0 (0.7, 1.6)	1.3 (0.7, 2.4)
	2	2.6 (2.1, 3.2)	4.2 (3.2, 5.5)	3.8 (2.9, 5.0)	7.5 (5.5, 10.3)	2.4 (1.7, 3.2)	2.9 (1.9, 4.4)
	3	6.5 (5.4, 7.7)	6.5 (4.9, 8.5)	9.8 (7.9, 12.2)	11.8 (8.6, 16.0)	5.8 (4.7, 7.2)	4.8 (3.3, 7.1)
	4	13.9 (11.9, 16.1)	13.6 (10.5, 17.5)	18.6 (15.1, 22.7)	15.5 (11.0, 21.5)	11.0 (9.0, 13.3)	9.8 (6.8, 13.8)
	5	25.9 (22.5, 29.6)	27.1 (21.3, 33.7)	37.4 (31.9, 43.3)	27.0 (19.9, 35.4)	15.5 (12.2, 19.5)	14.5 (10.0, 20.6)
	6	44.2 (38.6, 49.9)	32.4 (24.4, 41.6)	58.7 (51.1, 65.8)	22.8 (14.1, 34.5)	25.5 (20.2, 31.7)	15.1 (9.1, 24.2)
	7	50.7 (43.5, 57.9)	43.7 (32.8, 55.3)	73.9 (64.8, 81.2)	31.3 (19.5, 46.0)	33.0 (23.5, 44.1)	39.0 (27.7, 51.7)
	8	73.7 (62.8, 82.3)	69.4 (56.2, 80.0)	74.9 (62.6, 84.2)	51.0 (31.6, 70.2)	58.3 (44.6, 70.9)	
	9	83.2 (69.7, 91.4)	76.4 (56.0, 89.1)	87.6 (75.9, 94.1)	33.3 (18.6, 52.2)		
Extern- alising subscale score	AUC	0.89	0.86	0.90	0.81	0.81	0.77
	0	0.3 (0.1, 0.7)	0.4 (0.1, 1.2)	0.3 (0.2, 0.6)	0.5 (0.3, 1.0)	0.4 (0.1, 2.5)	0.0 -
	1	0.1 (0.0, 0.4)	0.5 (0.2, 1.4)	0.9 (0.6, 1.4)	2.1 (1.3, 3.5)	0.0 -	0.5 (0.1, 4.1)
	2	0.5 (0.2, 0.9)	0.8 (0.4, 1.6)	0.7 (0.4, 1.2)	1.9 (1.1, 3.1)	0.4 (0.1, 1.5)	0.4 (0.1, 2.9)
	3	0.7 (0.4, 1.1)	1.8 (1.1, 2.7)	2.0 (1.4, 2.9)	3.2 (2.1, 4.9)	0.6 (0.3, 1.6)	2.1 (0.9, 5.1)
	4	1.3 (0.9, 1.9)	1.9 (1.2, 3.1)	1.6 (1.0, 2.6)	2.9 (1.7, 4.9)	1.8 (1.1, 3.1)	1.7 (0.8, 3.5)
	5	2.1 (1.5, 2.9)	2.8 (1.8, 4.3)	2.3 (1.5, 3.5)	5.4 (3.6, 8.1)	3.0 (2.0, 4.3)	3.3 (1.9, 5.7)
	6	2.1 (1.5, 3.0)	3.5 (2.3, 5.3)	2.6 (1.7, 4.1)	6.3 (4.0, 9.9)	2.4 (1.6, 3.6)	1.6 (0.8, 3.4)
	7	2.7 (1.9, 3.8)	4.4 (2.9, 6.6)	5.1 (3.4, 7.6)	7.8 (5.2, 11.7)	4.0 (2.9, 5.6)	4.8 (2.9, 8.0)
	8	5.5 (4.2, 7.1)	5.7 (3.8, 8.4)	5.8 (4.1, 8.3)	7.3 (4.4, 11.8)	4.7 (3.3, 6.6)	5.8 (3.8, 8.9)
	9	7.5 (5.9, 9.6)	7.3 (5.1, 10.4)	9.1 (6.8, 12.3)	14.3 (9.5, 20.9)	7.4 (5.5, 9.9)	8.4 (5.4, 13.0)
	10	10.5 (8.2, 13.3)	13.5 (9.6, 18.6)	15.3 (11.8, 19.7)	9.2 (5.4, 15.0)	11.1 (8.3, 14.7)	9.1 (5.3, 15.2)
	11	15.6 (12.5, 19.2)	13.0 (8.6, 19.1)	19.3 (15.2, 24.2)	17.8 (11.8, 26.0)	17.8 (13.8, 22.6)	9.8 (5.7, 16.3)
	12	20.7 (17.1, 24.9)	16.2 (11.6, 22.1)	22.0 (17.3, 27.6)	18.8 (11.8, 28.8)	18.1 (13.6, 23.8)	12.8 (7.2, 21.9)
	13	32.9 (27.7, 38.6)	28.3 (21.0, 36.9)	34.2 (27.7, 41.3)	22.6 (14.8, 32.9)	24.2 (17.9, 31.8)	16.0 (9.0, 26.9)
	14	45.0 (38.3, 51.9)	36.8 (26.0, 49.0)	43.3 (35.9, 51.0)	17.5 (10.2, 28.2)	25.2 (17.1, 35.4)	10.3 (3.3, 27.7)
	15	62.4 (54.1, 70.1)	47.6 (35.0, 60.4)	53.3 (44.1, 62.4)	39.7 (26.9, 54.2)	33.5 (21.6, 47.8)	40.4 (27.2, 55.2)
	16	58.9 (49.4, 67.7)	60.1 (45.1, 73.4)	77.8 (67.1, 85.7)	32.4 (17.9, 51.2)	55.6 (42.0, 68.4)	
	17	65.3 (53.0, 75.8)	74.9 (61.9, 84.6)	81.2 (68.3, 89.7)	36.8 (24.5, 51.1)		
18	85.3 (70.5, 93.3)		77.2 (57.8, 89.3)				
19	94.1 (79.0, 98.5)		89.6 (72.2, 96.6)				
20							

AUC=area under the curve, determined by receiver operating characteristic analyses. High scores were grouped once the number of children per SDQ point fell to 20 or fewer to prevent very small numbers leading to instability in the prevalence estimates. The last prevalence estimate presented for each informant and at each time point therefore corresponds to the prevalence for that number of SDQ points or above.

Figure 4: Prevalence of behavioural disorder at baseline and follow-up by the parent, teacher and youth behavioural and externalising SDQ subscales at baseline



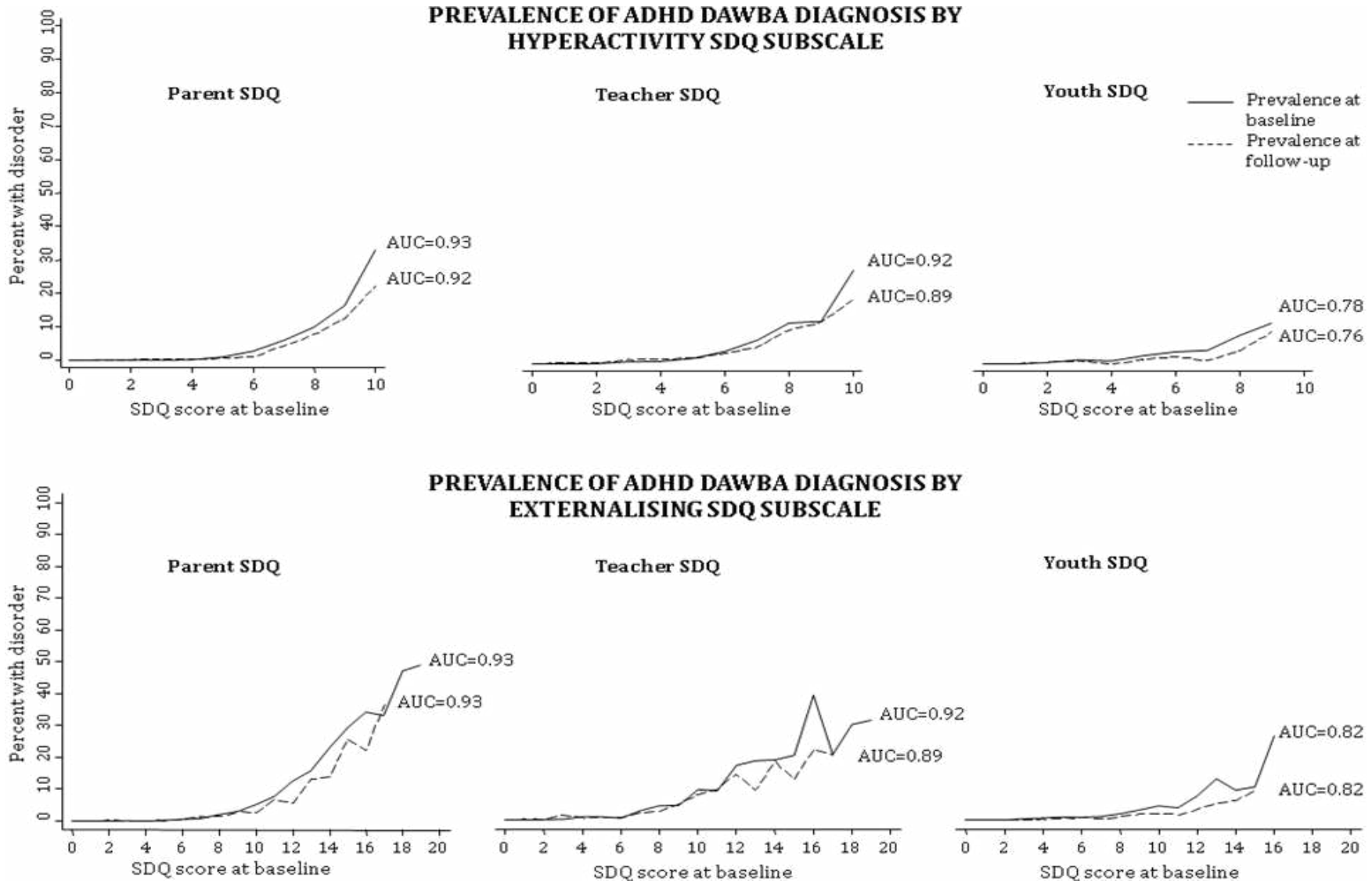
AUC=area under the curve

Table 9: Prevalence of ADHD at baseline and follow-up by the parent, teacher and youth hyperactivity and externalising SDQ subscales at baseline

		HYPERACTIVITY DISORDER					
		Parent SDQ		Teacher SDQ		Youth SDQ	
		Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI	Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI	Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI
N		18222	7901	14263	6247	7678	3408
Hyper-activity subscale score	AUC	0.93	0.92	0.92	0.89	0.78	0.76
	0	0.0 -	0.0 -	0.0 (0.0, 0.2)	0.0 -	0.0 -	0.0 -
	1	0.1 (0.0, 0.3)	0.1 (0.0, 0.8)	0.1 (0.1, 0.3)	0.3 (0.1, 1.0)	0.0 -	0.0 -
	2	0.0 (0.1, 0.3)	0.1 (0.0, 0.6)	0.2 (0.1, 0.5)	0.1 (0.0, 0.5)	0.5 (0.2, 1.1)	0.3 (0.1, 1.5)
	3	0.1 (0.0, 0.3)	0.2 (0.1, 0.8)	0.7 (0.4, 1.2)	1.2 (0.6, 2.4)	1.2 (0.7, 2.0)	0.8 (0.3, 2.0)
	4	0.2 (0.1, 0.5)	0.3 (0.1, 1.0)	0.9 (0.5, 1.7)	1.1 (0.5, 2.6)	0.9 (0.5, 1.5)	0.0 -
	5	1.0 (0.7, 1.6)	0.5 (0.2, 1.4)	1.8 (1.2, 2.8)	1.2 (0.7, 2.3)	2.5 (1.8, 3.5)	0.7 (0.3, 1.8)
	6	2.7 (2.0, 3.8)	1.0 (0.4, 2.3)	3.8 (2.7, 5.5)	3.0 (1.6, 5.6)	3.7 (2.5, 5.2)	1.7 (0.8, 3.6)
	7	5.9 (4.6, 7.7)	4.3 (2.7, 6.9)	7.2 (5.3, 9.8)	4.2 (2.2, 7.7)	4.1 (2.7, 6.4)	0.7 (0.1, 3.4)
	8	10.0 (7.9, 12.6)	7.6 (5.1, 11.3)	12.4 (9.7, 15.8)	8.6 (5.7, 12.9)	8.6 (5.9, 12.4)	3.8 (1.5, 9.7)
	9	16.4 (13.4, 20.0)	12.6 (8.8, 17.6)	12.8 (9.6, 16.9)	10.5 (6.4, 16.7)	13.7 (8.7, 21.0)	5.2 (2.0, 12.7)
10	33.0 (28.8, 37.4)	22.1 (16.4, 29.0)	28.1 (24.2, 32.4)	14.3 (10.2, 19.7)	8.5 (3.2, 21.0)		
Extern-alising subscale score	AUC	0.93	0.93	0.92	0.89	0.82	0.82
	0	0.0 -	0.0 -	0.03 (0.05, 0.2)	0.0 -	0.0 -	0.0 -
	1	0.0 -	0.0 -	0.1 (0.01, 0.4)	0.2 (0.0, 0.9)	0.0 -	0.0 -
	2	0.1 (0.0, 0.4)	0.1 (0.0, 1.0)	0.2 (0.1, 0.5)	0.1 (0.0, 0.5)	0.0 -	0.0 -
	3	0.1 (0.0, 0.4)	0.0 -	0.3 (0.1, 0.7)	1.5 (0.8, 3.0)	0.4 (0.1, 1.4)	0.0 -
	4	0.0 -	0.0 -	1.0 (0.6, 1.9)	0.7 (0.2, 2.1)	0.6 (0.2, 1.4)	0.3 (0.0, 1.9)
	5	0.3 (0.1, 0.7)	0.2 (0.0, 1.1)	1.0 (0.5, 1.9)	0.5 (0.1, 1.6)	1.0 (0.5, 1.9)	0.4 (0.1, 1.8)
	6	0.6 (0.3, 1.1)	0.4 (0.1, 1.3)	0.6 (0.2, 1.5)	0.6 (0.1, 2.8)	0.9 (0.5, 1.8)	0.7 (0.2, 2.3)
	7	0.8 (0.4, 1.6)	1.2 (0.5, 2.8)	3.0 (1.8, 4.8)	2.2 (1.1, 4.4)	1.1 (0.6, 2.2)	0.3 (0.0, 2.3)
	8	2.0 (1.3, 3.2)	1.3 (0.6, 2.9)	4.4 (2.9, 6.8)	2.0 (0.8, 5.0)	2.0 (1.1, 3.3)	0.9 (0.2, 3.0)
	9	3.1 (2.1, 4.5)	2.8 (1.5, 5.2)	4.6 (2.8, 7.5)	4.3 (2.1, 8.8)	3.1 (1.9, 5.1)	1.3 (0.4, 4.0)
	10	5.2 (3.6, 7.5)	2.7 (1.2, 5.9)	9.6 (6.7, 13.6)	4.7 (2.4, 9.1)	4.5 (2.8, 7.2)	1.2 (0.3, 4.3)
	11	7.7 (5.6, 10.6)	6.7 (3.9, 11.2)	9.1 (6.4, 12.9)	8.6 (4.8, 14.9)	3.9 (2.1, 7.2)	1.3 (0.3, 4.7)
	12	12.6 (9.6, 16.2)	5.5 (3.0, 10.0)	17.2 (12.9, 22.5)	11.6 (6.3, 20.2)	7.6 (4.7, 12.1)	2.4 (0.6, 8.6)
	13	15.9 (12.0, 20.7)	12.9 (8.0, 20.1)	18.6 (13.7, 24.8)	10.7 (5.2, 20.9)	13.1 (8.2, 20.2)	5.2 (1.8, 14.3)
	14	22.9 (17.6, 29.3)	13.8 (7.7, 23.6)	18.8 (13.5, 25.7)	18.2 (10.6, 29.5)	9.4 (4.7, 17.8)	5.1 (1.0, 22.0)
	15	29.4 (21.3, 39.1)	25.8 (15.6, 39.5)	20.3 (13.4, 29.6)	7.4 (3.4, 15.2)	10.5 (4.4, 23.0)	7.2 (2.5, 18.7)
	16	34.2 (25.0, 44.9)	22.1 (12.7, 35.7)	39.1 (28.3, 51.0)	17.6 (7.8, 34.9)	26.6 (16.0, 40.9)	
	17	33.2 (22.9, 45.5)	36.5 (25.7, 48.9)	20.6 (11.7, 33.7)	17.7 (9.6, 30.4)		
18	47.2 (31.5, 63.5)		29.9 (17.4, 46.3)				
19	49.1 (32.0, 66.4)		31.3 (16.3, 51.7)				
20							

AUC=area under the curve, determined by receiver operating characteristic analyses. High scores were grouped once the number of children per SDQ point fell to 20 or fewer to prevent very small numbers leading to instability in the prevalence estimates. The last prevalence estimate presented for each informant and at each time point therefore corresponds to the prevalence for that number of SDQ points or above.

Figure 5: Prevalence of ADHD at baseline and follow-up by the parent, teacher and youth hyperactivity and externalising SDQ subscales at baseline



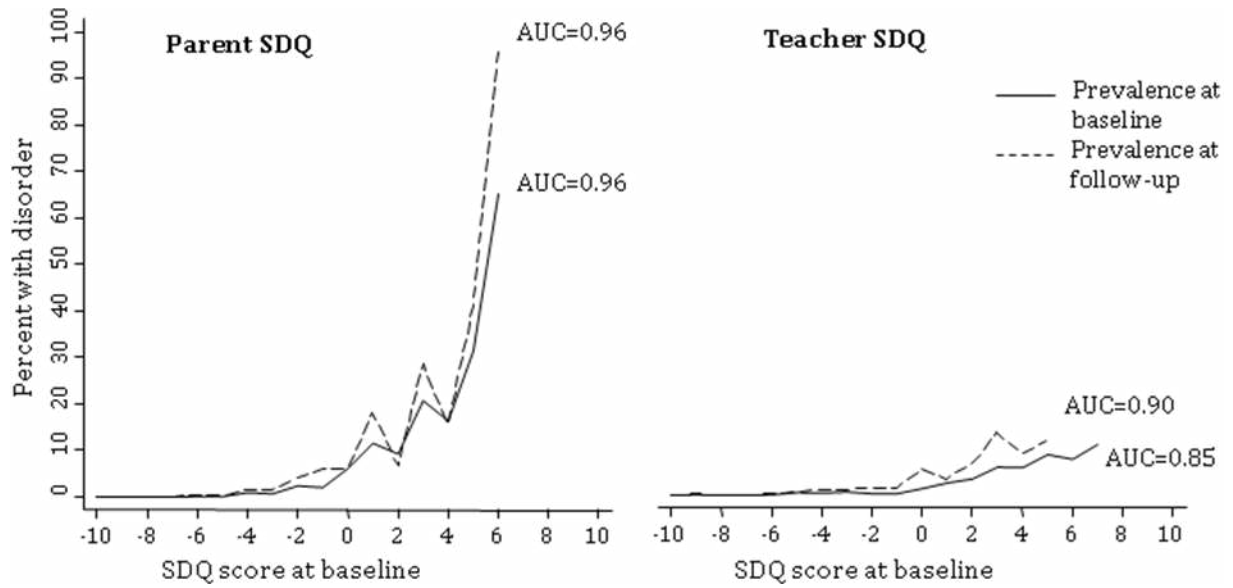
AUC=area under the curve

Table 10: Prevalence of ASD at baseline and follow-up by the parent, teacher and youth SDQ autism screening score (peer problems subscale minus prosocial subscale) at baseline

		AUTISM DISORDER			
		Parent SDQ		Teacher SDQ	
		Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI	Baseline prevalence & 95% CI	Follow-up prevalence & 95% CI
	N	18222	7901	14263	6247
Autism screening score	AUC	0.96	0.96	0.85	0.90
	-10	0.0 -	0.0 -	0.0 (0.0, 0.3)	0.0 -
	-9	0.0 -	0.0 -	0.1 (0.0, 0.5)	0.0 -
	-8	0.0 (0.0, 0.2)	0.0 (0.0, 0.2)	0.1 (0.0, 0.5)	0.0 (0.0, 0.2)
	-7	0.0 -	0.0 -	0.1 (0.0, 0.4)	0.0 -
	-6	0.1 (0.1, 0.4)	0.4 (0.1, 1.3)	0.1 (0.0, 0.6)	0.2 (0.1, 1.3)
	-5	0.2 (0.0, 0.7)	0.2 (0.0, 1.4)	0.6 (0.3, 1.2)	0.6 (0.0, 1.4)
	-4	1.0 (0.5, 2.0)	1.4 (0.6, 3.2)	0.5 (0.2, 1.1)	0.8 (0.6, 3.2)
	-3	0.8 (0.3, 2.0)	1.6 (0.6, 4.4)	0.7 (0.3, 1.5)	0.9 (0.6, 4.4)
	-2	2.4 (1.2, 4.7)	4.1 (1.9, 8.9)	0.4 (0.1, 1.3)	1.0 (1.9, 8.9)
	-1	2.1 (0.9, 4.9)	5.9 (2.7, 12.5)	0.4 (0.1, 1.6)	1.0 (2.7, 12.5)
	0	6.2 (3.3, 11.2)	6.0 (2.2, 15.5)	1.4 (0.6, 3.5)	4.7 (2.2, 15.5)
	1	11.6 (5.8, 21.8)	18.1 (7.6, 37.2)	2.7 (1.3, 5.6)	3.5 (7.6, 37.2)
	2	9.3 (3.9, 20.5)	6.6 (1.4, 25.9)	3.5 (1.6, 7.6)	6.1 (1.4, 25.9)
	3	20.7 (11.6, 34.3)	28.5 (13.9, 49.6)	6.0 (3.0, 11.6)	11.2 (13.9, 49.6)
	4	16.5 (7.2, 33.4)	15.5 (3.5, 48.4)	5.9 (2.5, 13.4)	10.0 (3.5, 48.4)
	5	31.5 (15.7, 53.1)	41.1 (17.9, 69.1)	8.8 (3.7, 19.5)	13.6 (17.9, 69.1)
6	65.3 (44.4, 81.6)	96.3 (76.8, 99.5)	7.8 (2.5, 21.6)		
7			10.9 (4.1, 25.9)		
8					
9					
10					

AUC=area under the curve, determined by receiver operating characteristic analyses. High scores were grouped once the number of children per SDQ point fell to 20 or fewer to prevent very small numbers leading to instability in the prevalence estimates. The last prevalence estimate presented for each informant and at each time point therefore corresponds to the prevalence for that number of SDQ points or above.

Figure 6: Prevalence of ASD at baseline and follow-up by the parent and teacher SDQ autism screening score (peer problems subscale minus prosocial subscale) at baseline



AUC=area under the curve

PART 2: Illustrations of the association of the SDQ subscales with DAWBA diagnoses at follow-up

Analyses come from models identical to those described in Table 5 of the main text, except that the SDQ subscales were entered as categorical terms by SDQ point rather than as linear scales. Subscale scores were also grouped once the number of children per point fell to 20 or fewer, to avoid estimates based on very small numbers. As in Table 5 of the main text, the prosocial score is reverse-scored to facilitate comparisons of effect sizes.

Figure 7: Independent association of the five parent SDQ subscales at baseline with DAWBA diagnoses at follow-up (note: this graph is also presented in the main text)

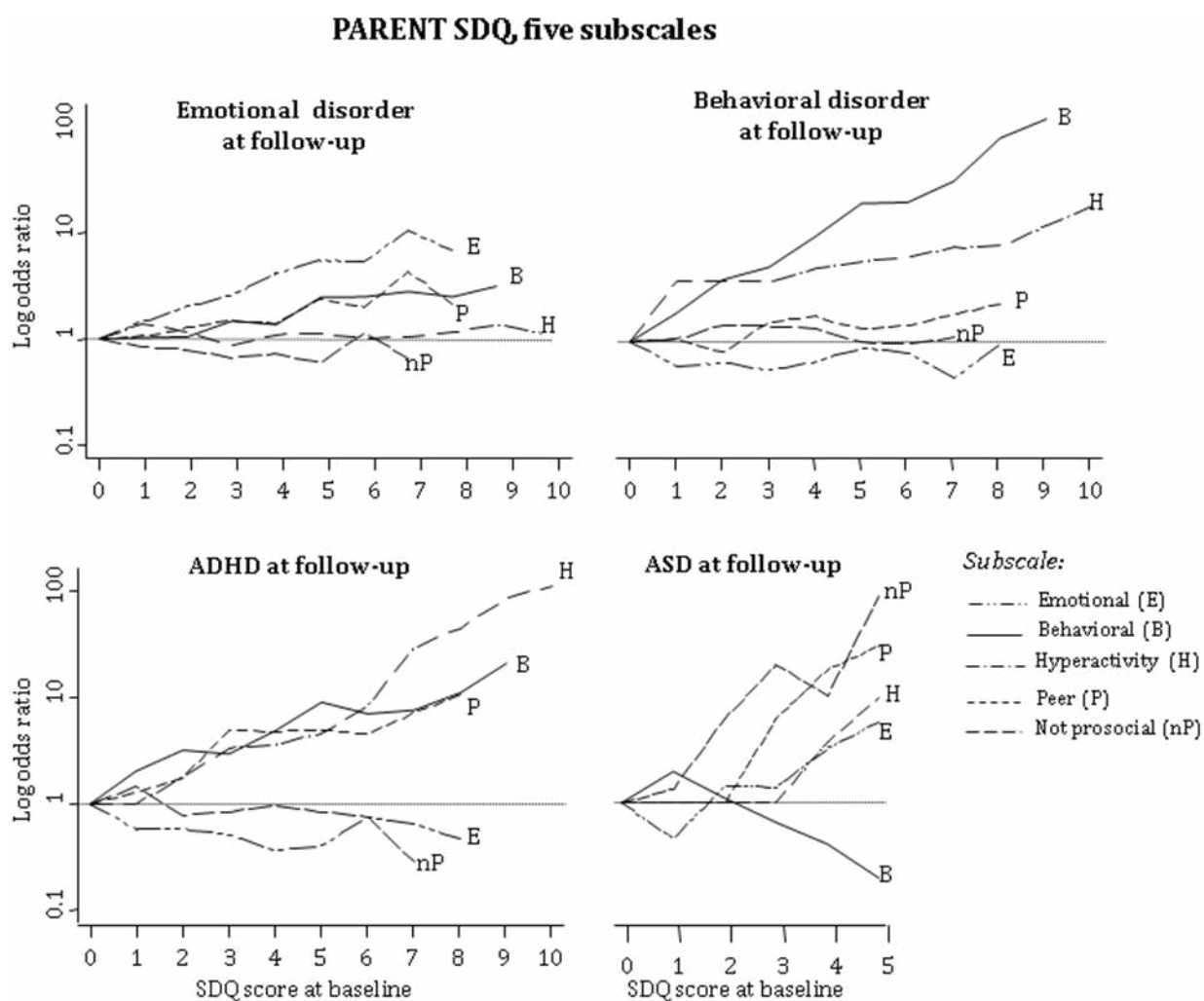


Figure 8: Independent association of the three parent SDQ subscales at baseline with DAWBA diagnoses at follow-up

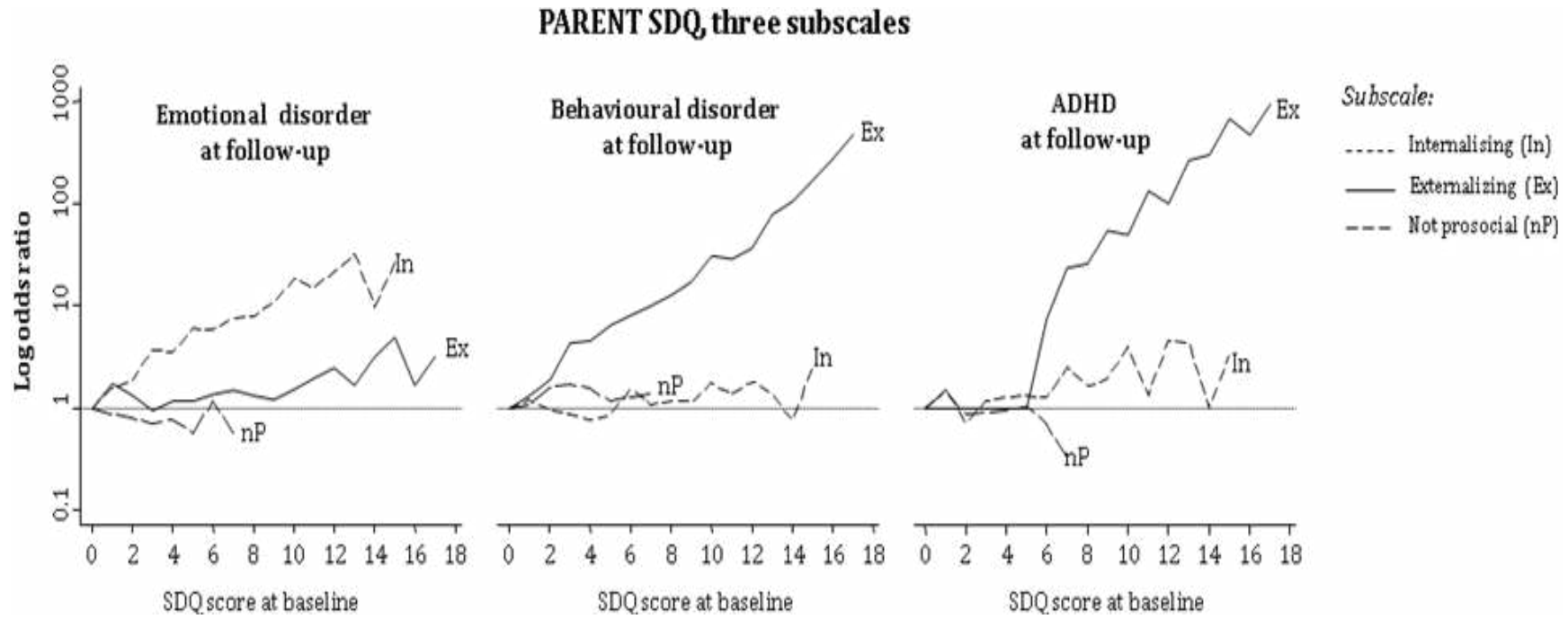


Figure 9: Independent association of the five teacher SDQ subscales at baseline with DAWBA diagnoses at follow-up

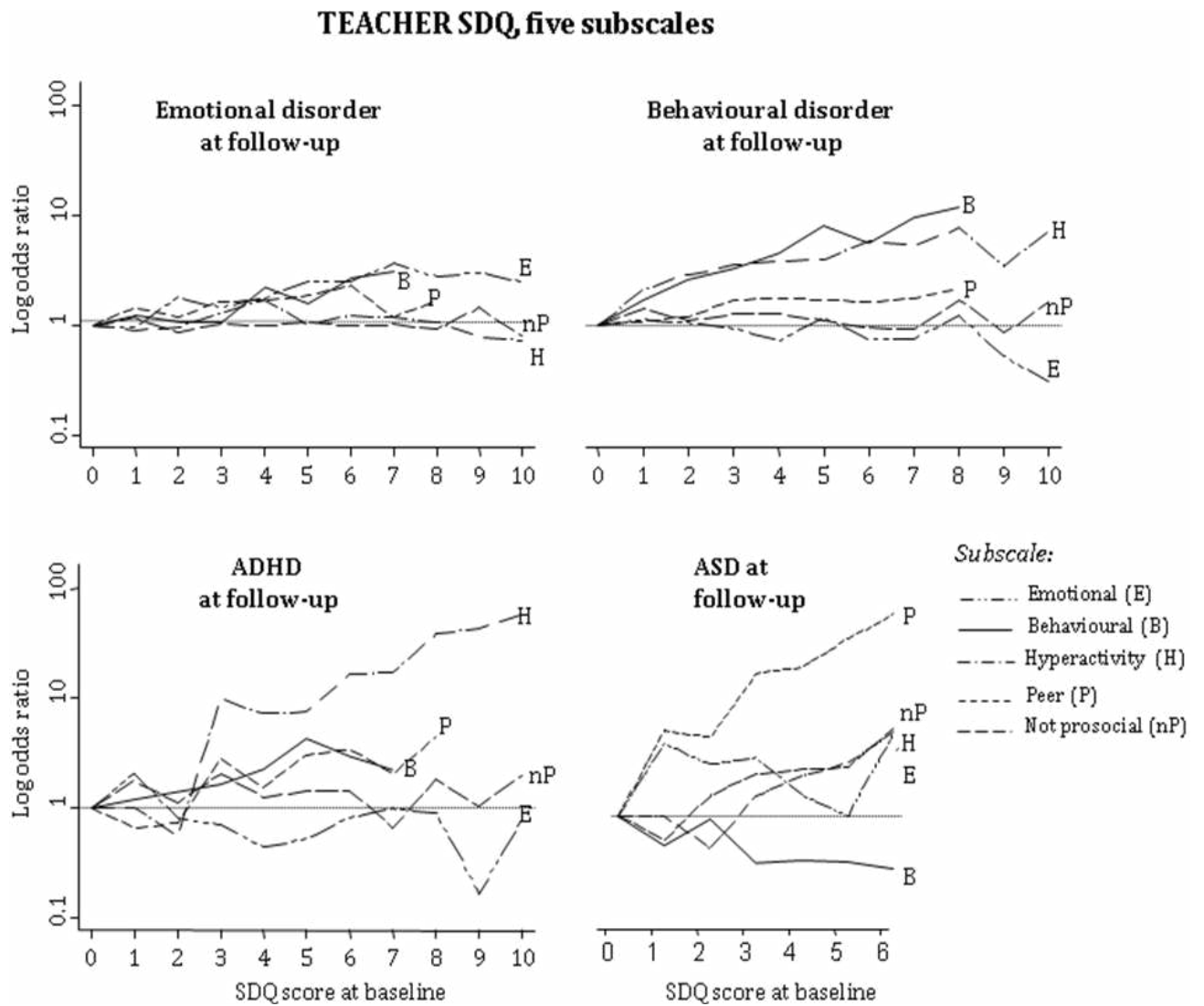


Figure 10: Independent association of the three teacher SDQ subscales at baseline with DAWBA diagnoses at follow-up

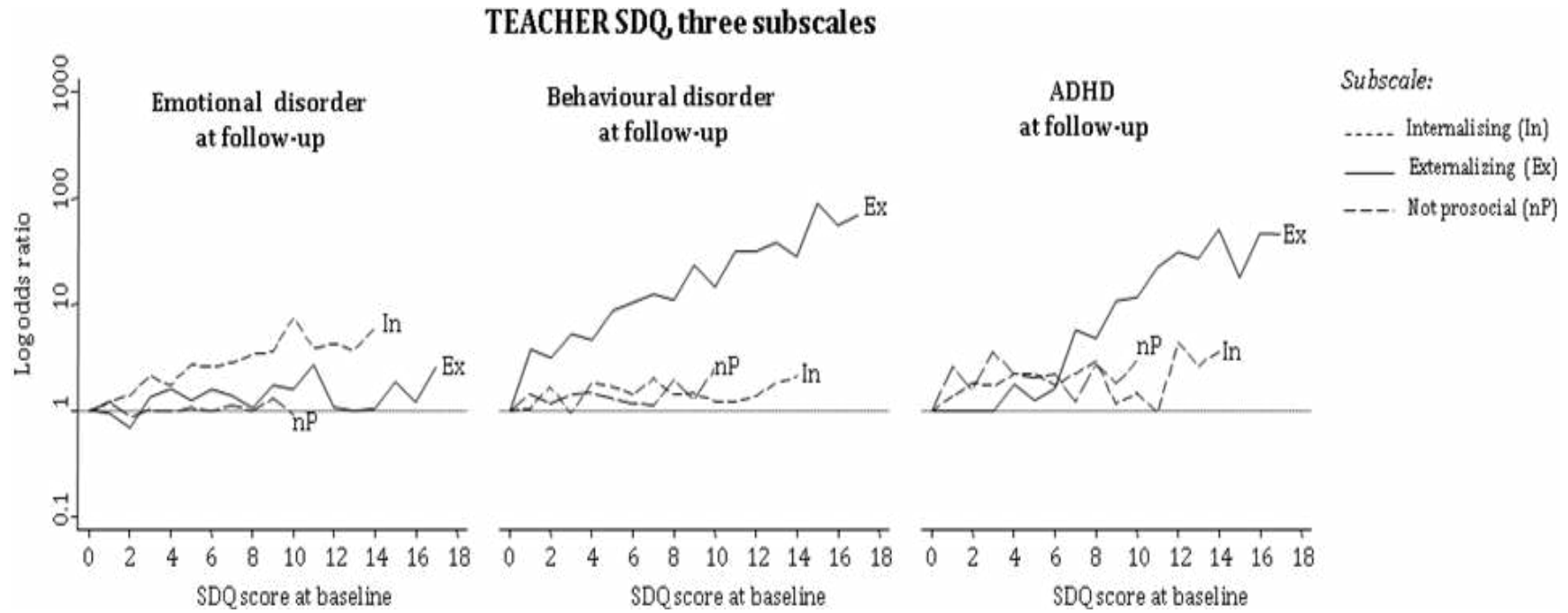


Figure 11: Independent association of the five youth SDQ subscales at baseline with DAWBA diagnoses at follow-up

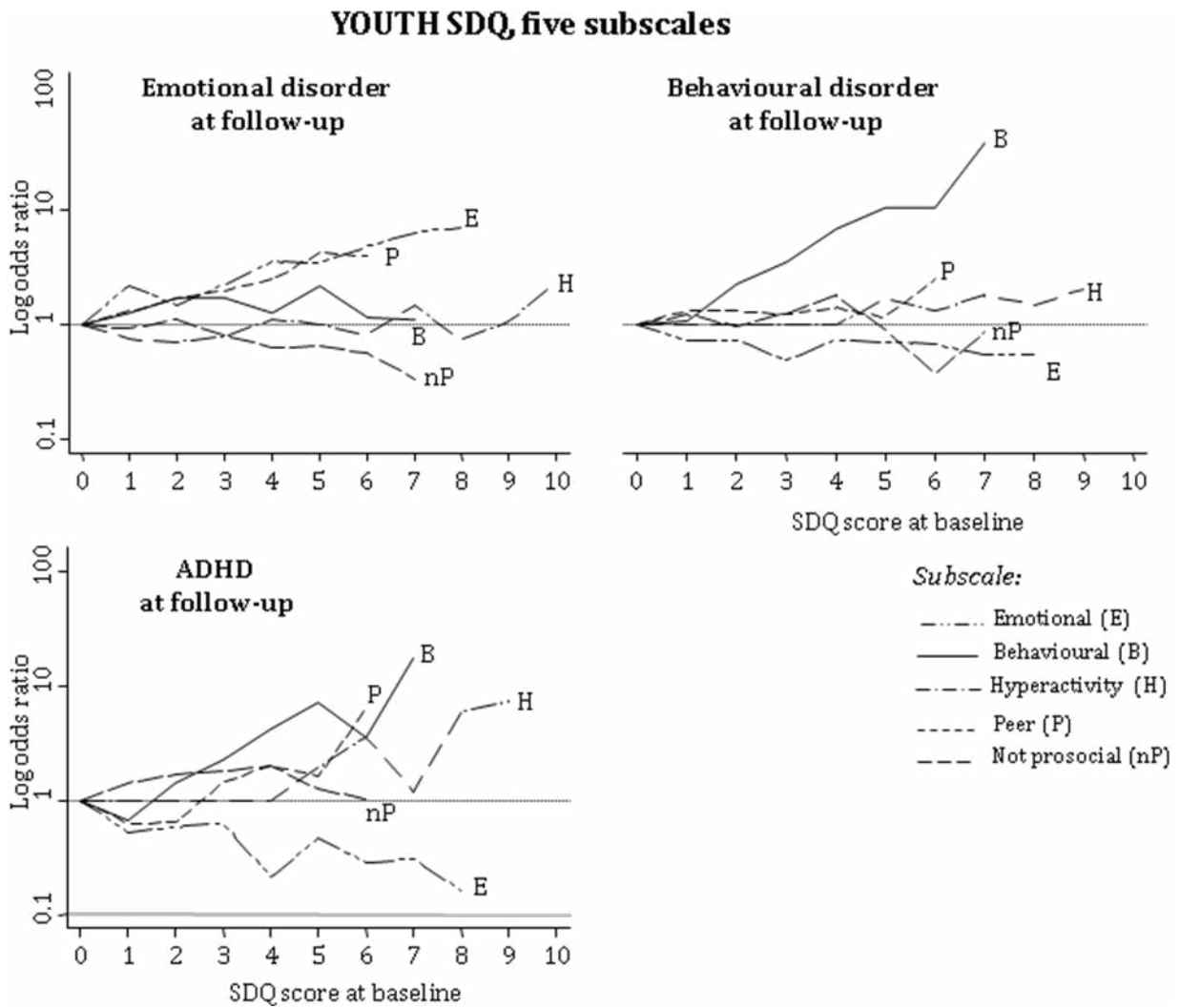


Figure 12: Independent association of the three youth SDQ subscales at baseline with DAWBA diagnoses at follow-up

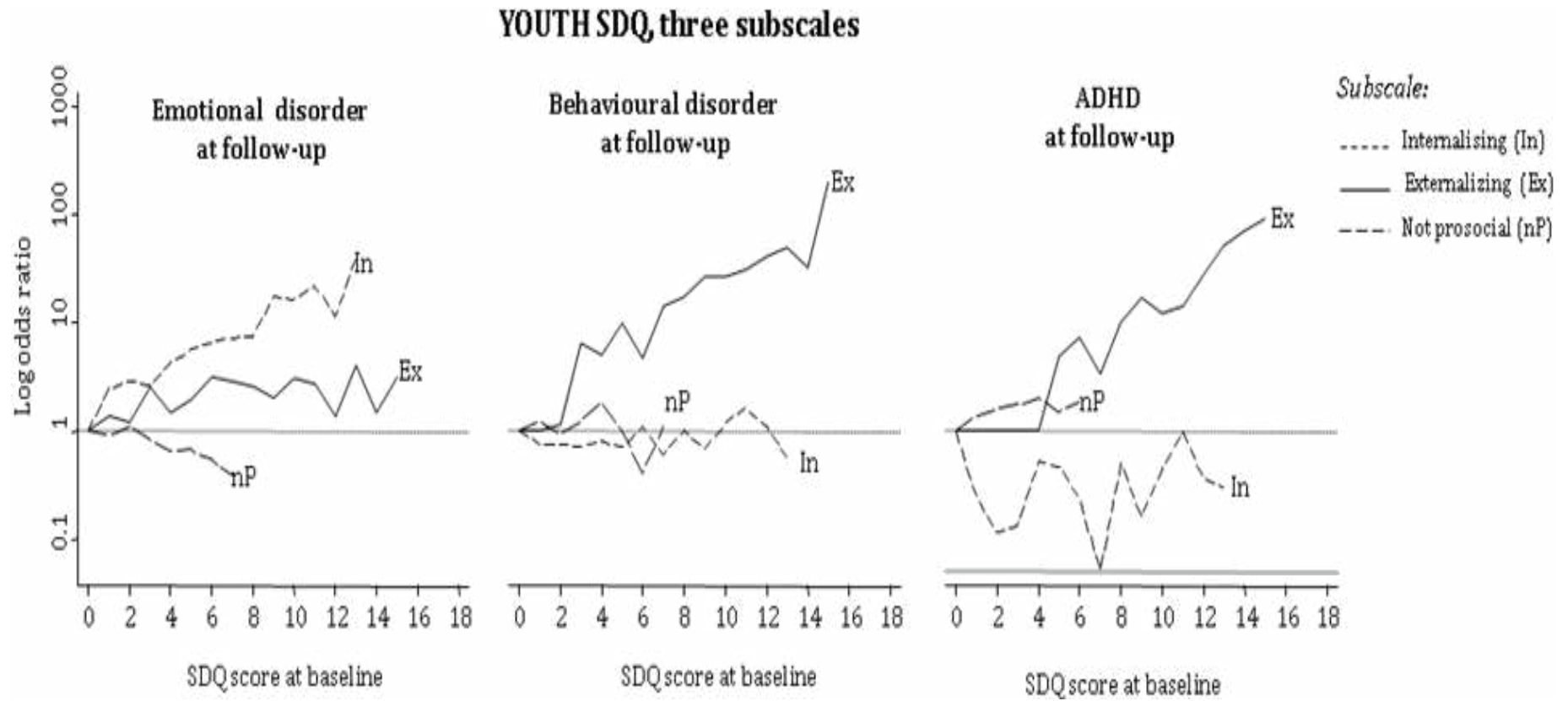


Figure 13: Independent association of the parent and teacher autism SDQ screening score at baseline with ASD DAWBA diagnoses at follow-up

