Title: A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire

1. Paula Kersten, Professor, Centre for Person Centred Research, Auckland University of Technology, Auckland, New Zealand
2. Karol J. Czuba, Research Officer, Centre for Person Centred Research, Auckland University of Technology, Auckland, New Zealand
3. Kathryn M. McPherson, Professor, Centre for Person Centred Research, Auckland University of Technology, Auckland, New Zealand
4. Margaret Dudley, Postdoctoral Research Fellow, Taupua Waiora Centre for Māori Health Research, Auckland University of Technology, Auckland, New Zealand
5. Hinemoa Elder, Professorial Fellow, Te Whare Mātai Aronui, Te Whare Wānanga o Awanuiārangi, Auckland, New Zealand
6. Robyn Tauroa, Research Assistant, Centre for Person Centred Research, Auckland University of Technology, Auckland, New Zealand
7. Alain C. Vandal, Department of Biostatistics and Epidemiology, Auckland University of Technology & Ko Awatea Health Intelligence and Informatics, Counties Manukau District Health Board, Auckland, New Zealand

Corresponding author:

Professor Paula Kersten, Centre for Person Centred Research, Auckland University of Technology, Private Bag 92006, Auckland, 1142, New Zealand. pkersten@aut.ac.nz. Tel +64 9 921 9180. Fax: +64 9 9219706.

Address where the work was carried out: Centre for Person Centred Research, Auckland University of Technology, Private Bag 92006, Auckland, 1142, New Zealand.

Abstract

This paper synthesised evidence for the validity and reliability of the Strengths and Difficulties Questionnaire in children aged 3-5. A systematic review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement guidelines was carried out. Study quality was rated using the Consensus-based Standards for the Selection of Health Measurement Instruments. 41 studies were included (56 manuscripts). Two studies examined content and cultural validity, revealing issues with some questions. Six studies discussed language validations with changes to some wording recommended. There was good evidence for discriminative validity (Area Under the Curve ≥ 0.80), convergent validity (weighted average correlation coefficients ≥0.50, except for the Prosocial scale), and the 5-factor structural validity. There was limited support for discriminant validity. Sensitivity was below 70% and specificity above 70% in most studies that examined this. Internal consistency of the total difficulty scale was good (weighted average Chronbach's alpha parents' and teachers' version 0.79 and 0.82) but weaker for other subscales (weighted average parents' and teachers' range 0.49-0.69 and 0.69-0.83). Inter-rater reliability between parents was moderate (correlation coefficients range 0.42-0.64) and between teachers strong (range 0.59-0.81). Cross-informant consistency was weak to moderate (weighted average correlation coefficients range 0.25-0.45). Test-retest reliability was mostly inadequate. In conclusion, the lack of evidence for cultural validity, criterion validity and test-retest reliability should be addressed given wide-spread implementation of the tool in routine clinical practice. The moderate level of consistency between different informants indicate that an assessment of a pre-schooler should not rely on a single informant.

Keywords

Strengths and Difficulties Questionnaire, SDQ, validity, reliability

Background

Behavioural and emotional problems in pre-schoolers can impact upon their transition into primary school (Eivers, Brendgen, & Borge, 2010; White, Connelly, Thompson, & Wilson, 2013), lead to on-going problems in middle-childhood (Kim-Cohen et al., 2009) and adulthood (Kim-Cohen et al., 2003), and affect educational achievement (Bierman et al., 2013). Behavioural problems in children as young as three have been shown to be predictive of problems later in life, including depression and anti-social personality disorders (Caspi, Moffitt, Newman, & Silva, 1996; Kessler et al., 2005). A key preventative strategy is therefore to enhance identification of children with behavioural problems from a pre-school age, so that support programmes can be put in place (Doughty, 2005).

Many countries use the Strengths and Difficulties Questionnaire for parents (SDQ-P) and for teachers (SDQ-T) to screen children (R. Goodman, 1997; R. Goodman, Meltzer, & Bailey, 1998). The SDQ is a 25-item questionnaire for assessing children's psychosocial attributes (positive and negative behaviours), made up five subscales: Emotional Symptoms, Conduct Problems, Hyperactivity, Peer Problems, and Prosocial Behaviour (R. Goodman, 1997; R. Goodman et al., 1998). Higher scores on the four subscales that report on difficulties reflect more significant problems, whereas higher scores on the Prosocial subscale denote better social behaviour. Scores from the first four subscales are summed to give an overall Difficulty score ranging from 0-40. Score distributions in large populations have been used to derive score thresholds for each subscale, as well as the total Difficulties score. These are used to classify children's difficulties as 'normal', 'borderline' and 'abnormal'. The SDQ also includes a page asking whether the reported difficulties cause the child distress (1 item) or impairment in their daily life (4 items for parents and youth, and 2 items for teachers) (R.

Goodman, 1999). Whilst answers to these questions are useful for clinicians they are not included in the scoring of the SDQ.

A recent review of 48 studies using the SDQ in 4-12 year olds concluded the SDQ was a good screening instrument but that further evidence for predictive validity in longitudinal studies was required (Stone, Otten, Engels, Vermulst, & Janssens, 2010). Our scoping indicated this review did not capture all relevant psychometric studies of the SDQ. In addition, whilst their review synthesised data they did not provide critical appraisal of the methodological quality of the included studies. Hence, we undertook a systematic review to identify and critically appraise evidence for the validity and reliability of the SDQ in pre-school children (aged 3-5).

Methods

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement guidelines (Liberati et al., 2009) and captured published studies reporting on reliability and/or validity of the SDQ-P/SDQ-T and that had included data on pre-school children (aged 3-5). Wildcards and truncation were used as specified by different databases:

("SDQ*" OR "strength* and difficult* questionnaire*") AND (psychometric* OR validat* OR validit* OR reliab* OR rasch* OR "factor* analysis*" OR "factor* structur*")

No date or language restrictions were set. The search included studies published up to 31st March 2014. Hand searches of reference lists of relevant articles were conducted. Studies that only included data on older children (aged 6 and above) were excluded. All references were downloaded into EndNote X4 (Thomson Reuters, 2010). Systematic review registers, such as

Cochrane (http://www.cochrane.org/) or PROSPERO (http://www.crd.york.ac.uk/prospero/) do not include reviews of outcome measures, hence this review protocol is not registered. Box 1 presents established definitions for psychometric properties, which we applied to all papers.

Two reviewers (KC, PK) screened all titles and abstracts and if needed the full article, to determine whether the article was eligible; any discrepancies were discussed until a consensus was reached. All studies were critically appraised by two reviewers (KC, PK). Psychometric properties reported in the studies were rated using the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) quality score (excellent, good, fair, poor) by obtaining the lowest rating of any item in a box (i.e. ‘‘worst score counts’’) (Mokkink et al., 2010; Terwee et al., 2012). For this review we considered inter-rater reliability as having been assessed when the study had evaluated consistency of scores between the same type of informants (e.g. two teachers). Studies that examined consistency between different types of informants (e.g. parent and teacher) who would be using different information on the child to derive their scores were considered to have examined cross-informant consistency (R. Goodman, 1997). If the paper did not explicitly state or report what had been done, we rated this as not having been done.

Quantitative data extraction followed COSMIN guidance and included data on study procedures, participants, assessments (if relevant), key findings from the appraisal, and reviewers’ comments. As with the critical appraisal, KC independently extracted data with a random sample of studies audited by the second reviewer (PK). Any uncertainties or discrepancies were discussed between the two reviewers and resolved by that discussion. Results from individual studies reported in multiple papers were combined to avoid risk of

bias across studies. For the purpose of rating whether reported results for each psychometric property were acceptable, we used the following criteria (Dancey & Reidy, 2007; Hu & Bentler, 1999; Streiner & Norman, 2008; Terwee et al., 2007):

- Cronbach's alpha ≥ 0.7 (for group use) and ≥ 0.85 (for use with individuals);
- Intraclass Correlation Coefficient (ICC) ≥ 0.70;
- Correlation coefficients for reliability ≥ 0.80;
- Correlation coefficients for convergent validity ≥ 0.50;
- Receiver Operating Characteristic (ROC) curves - Area Under the Curve (AUC) ≥ 0.80;
- Kappa coefficient ≥ 0.70;
- Sensitivity ≥ 80% and Specificity ≥ 60%;
- Confirmatory Factor Analysis (CFA): Root Mean Squared Error of Approximation (RMSEA) < 0.06 good fit, ≤0.08 acceptable fit; Comparative Fit Index (CFI), Goodness of Fit Index (GFI), (non) Normed Fit Index (NFI) and Tucker-Lewis Index (TLI) >0.90 good fit, 0.8-0.9 acceptable fit; (Standardised) Root Mean Squared Residual (SRMS) <0.08 good fit. Principal component or exploratory factor analyses were not included in the review since we were particularly interested in evaluating evidence for the established Goodman factor structure.

Two qualitative papers were identified that explored content and cultural validity. They were critiqued using the Critical Appraisal Skills Programme tool for qualitative studies (Critical Appraisal Skills Programme (CASP), 2003) and summarised narratively.

Data synthesis

Data from papers were extracted for each psychometric property. Sample size-weighted averages and standard deviations were calculated for internal consistency (Cronbach's alpha), cross-informant consistency (correlation coefficients) and convergent validity (correlations coefficients). Weighted standard deviations account only for between-study variability, as within-study standard errors were often not reported or not obtainable, notably in the case of Cronbach's alpha. All correlation coefficients types were considered equivalent for the purpose of computing summary statistics. Due to the possible heterogeneity of the groups and the different types of sample correlations involved, weighted summaries should be taken as indicative only.

Results

Fifty-six manuscripts were included reporting on 41 studies from 28 countries (Figure 1) with data from general *or* clinical populations (34 versus 13 manuscripts respectively; eight including both general and clinical samples) and one paper reviewing the translation of SDQ without any reference to a specific population. As noted above, definitions of psychometric terms are provided in Box 1.

Content validity

Studies were considered to have addressed content validity if they explicitely examined the degree to which the content of an instrument is an adequate reflection of the construct to be measured (Mokkink et al., 2010). This entailed questions from the COSMIN checklist such as relevance of the questions to the construct, to the study population and for the purpose of the measurement instrument. Two studies were included. Williamson et al. (2010) carried out a study in Aboriginal community-controlled health services. Participants included Aboriginal

parents, research assistants, youth workers, medical services staff and education officers. The study's limitations included a lack of detail around sampling, data saturation and an absence of participants' quotes to substantiate interpretation. Participants reported that the use of a questionnaire as opposed to a general conversation or interview was deemed culturally inappropriate and problematic for those with literacy issues. Inter-relationships with peers were considered of less importance than relationships with family and participants felt that many important aspects of children's behaviour and emotions were not covered by the SDQ. They also reported that the SDQ might not be completed honestly for fear of use of the data by other services or answers reflecting badly on their parenting skills. Participants recommended the re-wording of several questions and response scales, for example to enhance cultural clarity, although no questions were considered offensive.

White et al. (2013) included child development officers with direct responsibility for groups of children and head teachers from pre-school establishments in Scotland. This study also did not describe data saturation but otherwise provided a complete description of its methods. Despite the age of the children that staff worked with in this study (3 to 5), the SDQ used was the 4-16 year old version rather than the 3-4 year old version. Participants reported that using the SDQ provided a valuable opportunity to reflect on the emotional and social development of the children and to be able to share this with parents and the primary school. However, teachers reported that in most cases the SDQ did not reveal anything that they were not already aware of. Whilst most of the items were considered straightforward, participants reported that two items caused unease (often lies and cheats; steals from home, school or elsewhere). Staff expressed some concerns about parent reactions to the teacher-completed SDQ, about the SDQ leading to labelling of children. They also reported that the answers

would be dependent on who completes the tool, which is why some completed it as a team. In addition, the use of the tool was perceived as a paperwork burden.

Construct validity

Discriminative validity

Discriminative validity (mostly between clinical and community groups) was evaluated in 12 studies (Table 1). Overall, the quality of the studies was fair. Nine studies reported ROC curves and AUC values (Table 1, median sample size 338, range 94-845). These were acceptable in eight studies for SDQ-P and in four studies for SDQ-T (Table 1). The remaining three studies used different analytical approaches (distributional statistics, chi-squared test, kappa statistics and discriminant analysis) and supported the discriminative validity of SDQ (De Giacomo et al., 2012; R. Goodman, 1999; Petermann, Petermann, & Schreyer, 2010).

Convergent validity

Twenty-one studies were identified as having examined convergent validity with 17 being included in this review. Two (R. Goodman, 1999; Holtmann, Becker, Banaschewski, Rothenberger, & Roessner, 2011) were excluded as they did not evaluate the original 5-factor structure; the sample size for one was too small according to the COSMIN criteria (n=48) (Gruenert, Ratnam, & Tsantefski, 2006); and one did not meet our pre-defined criteria (Hill & Hughes, 2007) for convergent validity set out in Box 1 (i.e. they looked at convergence of the scores between parent and teacher versions of SDQ, rather than between SDQ and another measure). Sixteen studies used correlation coefficients and were included in data synthesis (Table 2). The one remaining study (Bourdon, Goodman, Rae, Simpson, & Koretz, 2005) used logistic regression and reported a significant association ($p < .0001$) between

children's service use and parent-rated Total Difficulty score, i.e. 45% of children with high Total difficulties score (above 90th percentile) used at least one type of mental health services. However, we cannot be certain if non-use of services is due to the validity of the SDQ or unavailability or non-uptake of services.

The methodological quality of the majority of the 16 included studies (median sample size n=182, range 21-1940) was fair according to the COSMIN rating (Supplementary file A). Apart from three exceptions (R. Goodman, 1997; Hawes & Dadds, 2004; C. L. Mieloo et al., 2014), all studies reported moderate or strong correlation coefficients. However, the coefficients reported on the Prosocial subscale were low in magnitude in 7 out of 8 studies. Weighted average correlation coefficients indicate that convergent validity of SDQ is acceptable for the Total Difficulties (Parent 0.67, Teacher 0.78), Emotional, Conduct and Hyperactivity subscales (Parent 0.55-0.63, Teacher 0.54-0.80) but unacceptable for the Peer Problems (0.49 for both SDQ versions) and Prosocial (Parent 0.18, Teacher 0.35) scales (Table 2).

Discriminant validity

Eight studies reported having examined discriminant validity. However, six evaluated the ability of the SDQ to differentiate between extreme groups of respondents, i.e. discriminative validity rather than discriminant validity (see definitions in Box 1), and they are reported under discriminative validity section. The two included studies used a Multitrait-multimethod (MTMM) analysis (A. Goodman, Lamping, & Ploubidis, 2010; Hill & Hughes, 2007) comparing scores between dissimilar subscales of the SDQ. Both studies were of fair quality according to COSMIN criteria and reported limited support for the discriminant validity of the SDQ scales.

Structural validity

Twenty-seven studies evaluated the structural validity of the SDQ as specified by Goodman, of which 17 used a CFA and were included (Table 3 and Supplementary file B, median sample size 1068, range 129-56864). One study carried out a CFA on each of the SDQ scales and is therefore not included in Table 3 (Thabet, Stretch, & Vostanis, 2000). Most studies were of fair quality with the most common weakness not describing how missing data were handled. One study dichotomised the SDQ scores (i.e. grouping categories 1 ‘somewhat true’ and 2 ‘certainly true’). All of the 13 studies with parents and the study with custodial grandparents demonstrated acceptable to good evidence for the 5-factor structure. Of the nine studies that examined structural validity of the teachers’ version of the SDQ, eight reported it as acceptable to good.

Cultural validity

Six studies discussed the translation process utilised for the SDQ into Arabic, Maltese, Bangla, Urdu and Chinese (quality ratings fair to excellent). Four of these mentioned that forward and backward translations were used, but insufficient detail of the process or changes made to translations during the process were provided (Alyahri & Goodman, 2006; Cefaia, Camillerib, Cooperc, & Saidd, 2011; R. Goodman, Renfrew, & Mullick, 2000; Thabet et al., 2000). Parents in the studies in Pakistan and the Gaza strip reported difficulties with literacy and required support to complete the Urdu and Arabic SDQ (Samad, Hollis, Prince, & Goodman, 2005; Thabet et al., 2000). Samad et al. (2005) provided specific examples that outlined the need for cross-cultural adaptations to some questions. For example, they translated words such as ‘steals’ and ‘lies’ more “subtly” but did not specify how the wording was changed.

Toh, Chow, Ting and Sewell (2008) raised concerns about the Chinese version of the SDQ and undertook an independent back-translation as recommended in the literature (Beaton, Bombardier, Guillemin, & Ferraz, 2000). They concluded problems with the Chinese version in use (Du, Kou, & Coghill, 2008), including: flow and grammar; wrongly written Chinese characters; some deviation in translation from the original meaning; problems with translation of the response category 'true'; additions of the verbs 'will' and 'can' that may change the meaning of the statement; and use of the same questionnaire for all age groups. However, as yet these changes have not been included in a revised version.

One study examined measurement invariance with respect to ethnicity between British Indian and British white children using data from the 1999 and 2004 British Child and Adolescent Mental Health Surveys (A. Goodman, Patel, & Leon, 2010). All parents completed the English version of the SDQ and the multi-group confirmatory factor analyses provided evidence of acceptable fit to the parent and teacher SDQ across ethnicity. One qualitative study addressed cultural validity and content validity of the SDQ for Aboriginals in Australia (Williamson et al., 2010) as previously discussed.

Criterion (concurrent and predictive) validity

Six studies (median sample size of 500, range 86-7984) examined criterion validity by comparing scores from the SDQ total difficulties and / or subscales to a 'gold standard' clinical diagnostic interview with clinical samples (Bekker, Bruck, & Sciberras, 2013; R. Goodman et al., 2000), community samples (Ezpeleta, Granero, la Osa, Penelo, & Domènech, 2012; R. Goodman, 2001; Mathai, Anderson, & Bourne, 2004) and children in care (R. Goodman, Ford, Corbin, & Meltzer, 2004). Methodological quality was fair in most cases.

Four studies reported sensitivity that was considered inadequate by our criteria (<70%) (Bekker et al., 2013; Ezpeleta et al., 2012; R. Goodman, 2001; Mathai et al., 2004). One study reported sensitivity of 63% for "private household children" as rated by their parents, but 85% for "looked-after children" (i.e. children at foster homes or residential homes) as rated by their carers (R. Goodman et al., 2004). Goodman et al. (2000) reported high sensitivity (>80%) of three SDQ subscale scores (Conduct, Emotional, Hyperactivity) in identifying children who were clinically diagnosed with a disorder. This study was carried out with children referred to a multidisciplinary child mental health clinic rather than a general population.

Most studies reported adequate specificity (>70%). One study showed inadequate specificity (47%) of the conduct subscale for their London sample (although a large number of the false negatives were children with possible conduct problems (R. Goodman et al., 2000)). Another study (Bekker et al., 2013) showed specificity was below 50% for both Emotional and Conduct subscales, resulting in relatively large numbers of children incorrectly being identified as having problems in this area.

Two studies used coefficients of determination or R^2 to identify the proportion of variance. Goodman and Goodman (2011) reported R^2=0.95 and R^2=0.91 for the parent and teacher versions, respectively. Goodman et al. (2012) analysed population SDQ data from seven countries and compared SDQ 'caseness' (prevalence based on the mean total difficulty scores, adjusted for the population's age and sex composition) against the measured prevalence of disorder using the Development and Well-being Assessment (DAWBA) tool. They reported average R^2=0.29 and R^2=0.56 for the parent and teacher versions, respectively.

The authors concluded that SDQ scores cannot be compared cross-nationally without population-specific norms.

Two studies used odds ratios to estimate the likelihood of receiving a diagnosis at baseline (concurrent validity) and three years later (predictive validity) (A. Goodman & Goodman, 2009; A. Goodman, Lamping, et al., 2010). Their findings generally supported criterion validity.

Internal consistency

Thirty-four manuscripts examined the internal consistency of the SDQ (median sample size of 739, range 48-22108). Five were not included in data synthesis: two did not look at the original 5-factor structure (Holtmann et al., 2011; Stringaris & Goodman, 2013); one reported scores combined across subscales (Niclasen, Teasdale, et al., 2012) and one across samples (Thabet et al., 2000); with sample size for one being too small (n=48) (Gruenert et al., 2006). We extracted 282 Cronbach's alphas from 26 studies (Table 4 and Supplementary file C). Of these, 150 (53.1%) fell above the acceptable threshold of 0.70, but only 16 (5.6%) were ≥ .85. The weighted average Cronbach's alpha for the SDQ-P total score was 0.79 and for the subscales it ranged between 0.49 and 0.69. Cronbach's alpha for SDQ-T (the teacher version of the scale) total score was 0.82, and for the subscales it ranged between 0.69 and 0.83. All subscales of the SDQ-P (the parent version of the scale) fell below the threshold of ≥ 0.70, which could be seen as an indication of inadequate internal consistency of those subscales. In general, the SDQ-T appears to have a higher internal consistency than the SDQ-P, and no single subscale presented Cronbach's alpha values acceptable for individual use, i.e. ≥ 0.85. Three studies used other statistics, specifically omega coefficients (Ezpeleta et al., 2012), model- based reliabilities (Gómez-Beneyto et al., 2013), and composite reliability (CR) and

average variance extracted (AVE) (Niclasen, Skovgaard, Andersen, Sømhovd, & Obel, 2012). Their findings supported the internal consistency of SDQ-P and SDQ-T.

Reliability

Inter-rater reliability

One study examined inter-rater reliability between two parents and between two teachers, using Spearman Ranked correlation coefficients (Borg, Pälvi, Raili, Matti, & Tuula, 2012). These ranged between 0.42 and 0.64 when parents' scores were compared and between 0.59 and 0.81 when teachers' scores were compared.

Cross-information consistency

Thirteen studies assessed consistency of scores between different types of informants. Of these, two reported only the range of correlation coefficients ((Birkás, Lakatos, Tóth, & Gervai, 2008): 0.31-0.65; (Cefaia et al., 2011): 0.14-0.37). Eleven studies were included in data synthesis (median sample size 512, range 99-7313) (Table 5). The quality of the studies was mostly fair. Correlation coefficients were weak to moderate, with weighted averages ranging between 0.25 and 0.45.

Test-retest reliability

Six studies assessed test-retest reliability of the SDQ (sample size median 592, range 34-2091, Table 6). In most cases, the methodological quality of the studies was fair. Most of the reported correlation coefficients indicate inadequate test-retest reliability. Only one study reported adequate test-retest reliability of the SDQ-P Total Difficulties score (*ICC*=.85) (R. Goodman, 1999). As for SDQ-T, stability of Total Difficulties and Hyperactivity-Inattention

scores was adequate in the one included study (.80 and .82, respectively) (R. Goodman, 2001).

Responsiveness

One study examined the responsiveness of the SDQ following services from a community child and adolescent mental health services (Mathai, Anderson, & Bourne, 2003). Quality of the study was fair with the main limitations being a lack of clarity on what intervention occurred during the 6-month period, the large loss to follow-up (66%), and a lack of *a priori* hypotheses. Improvements were observed on the SDQ total difficulty scale (effect size [ES] 0.45), emotional scale (ES 0.47) and conduct scale (ES 0.35), which concurred with changes on the clinician-administered Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA). However, it is not known if the findings would be similar for those who did not return to the service.

Discussion

This systematic review is the first one to use a standardised critical appraisal tool for the evaluation of psychometric properties of the SDQ. In addition to supporting evidence for a number of these psychometric properties we found lack of evidence for the test-retest reliability, cultural validity and criterion validity of the tool as well as poor convergent validity of the Peer problem and Prosocial scales. The lack of evidence is of concern given the tool is used across the world to identify which children need support to manage their behavioural and emotional problems. Given that such problems can impact upon their transition into primary school and affect educational achievement (Bierman et al., 2013;

Eivers et al., 2010; White et al., 2013) reliance on SDQ scores is inadequate to identify such children.

Evidence for the discriminative validity of the SDQ was good, in other words the SDQ is able to separate out populations hypothesised to have markedly different scores. The 5-factor structural validity of the SDQ was also good, providing confidence in the Goodman structure (R. Goodman, 2001) that tends to be employed in clinical practice including in New Zealand (Ministry of Health, 2008). Most studies demonstrated good evidence for the scale's convergent validity when compared with other scales measuring similar constructs, except for the Peer problem and Prosocial scale with this requiring further investigation.

Given the widespread use of the SDQ worldwide, we were surprised to find only one study that examined the content and cultural validity from the parents' perspectives (Williamson et al., 2010). This study identified concerns about the SDQ (e.g. fear of use of the data) and made recommendations for the re-wording of several questions and the response scales that would improve cultural clarity. However, there have as yet been no changes in the Australian version of the tool. Similarly, only one study was found that explicitly examined content validity as perceived by teachers (White et al., 2013). Further work on cultural validity therefore seems warranted.

At time of writing there are 79 different language versions available from the *Youth in Mind* website (http://www.sdqinfo.org/). Translations and adaptations are not permitted without the involvement of that study team, which provides confidence in the robustness of translations. Their procedures include forward and backward translations by teams of people and the final version is signed off by *Youth in Mind* (personal communication, www.youthinmind.info).

However, these data are not publicly available and hence we identified few studies explicitly describing the language translation or cross-cultural adaptation procedures. It was of concern that significant problems have been identified with the Chinese version (Toh et al., 2008) although the *Youth in Mind* group reported this version is currently being revised (personal communication, www.youthinmind.info).

Evidence for criterion validity of the SDQ was stronger in clinical than general population samples. Whilst this is perhaps unsurprising, it is of concern given the tool is specifically used in screening children to identify those who would benefit from further assessment or services. In addition, one study that pooled data from seven countries showed that the prevalence estimators derived from the SDQ scores spread very broadly across the countries (A. Goodman et al., 2012). Consequently, SDQ scores cannot be compared cross-nationally without population-specific norms.

Findings regarding the internal consistency of the SDQ Difficulty scale indicate that it is acceptable for comparing groups, but not adequate for clinical decision making. In addition, the internal consistency of the 5 subscales was not borne out in the review.

Goodman argued in his 1997 paper that parents and teachers make SDQ ratings based on different sources of information and that comparing their scores is therefore an investigation of cross-informant consistency as opposed to inter-rater reliability (R. Goodman, 1997). This view was echoed by Stone and colleagues (Stone et al., 2010), drawing on research by others (Achenbach, McConaughy, & Howell, 1987), who emphasise that informants such as parents or teachers see the children in different contexts and interact with the children in different ways. For this reason, Goodman recommends the use of correlation coefficients, rather than intra-class correlation coefficients and this has been followed by researchers examining the

SDQ (R. Goodman, 1997). Our weighted averages of coefficients between different informants ranged from 0.24 to 0.45 and were similar to those found by (Stone et al., 2010) (range 0.26 to 0.47). These authors claim a meta-analytical correlation coefficient of 0.27 between parents and teachers can be used as a benchmark of agreement or data quality and that therefore the weighted average in their review demonstrate good inter-rater agreement. We contend that the coefficient values found in Stone et al. (2010) and our review are actually rather low and indicate that at best 22% of variance can be explained by scores from different informants.

Our review has a number of strengths, including the use of a systematic and replicable search strategy, use of the PRISMA guidelines, two reviewers and a validated critical appraisal tool. In addition, we explicitly identified *a priory* the criteria by which we would judge if statistical findings were acceptable for each psychometric property. It is possible that we identified a larger number of study limitations of studies than others have (Stone et al., 2010) as the COSMIN tool is relatively new (Mokkink et al., 2010; Terwee et al., 2012). In addition, many included studies presented data on children with a wider age range than what we were particularly interested. Comparing the findings from the smaller number of studies which had only included younger children with the remaining studies suggest similar findings. However, further validation work in the younger age group seems warranted based on this review.

Conclusion

The systematic review has shown that the evidence for the discriminative and structural validity of the SDQ is strong, as is the evidence for convergent validity (apart from the Peer problems and Prosocial scales) and the internal consistency of the SDQ Total Difficulty scale.

The lack of evidence for other psychometric properties, in particular test-retest reliability, cultural validity and criterion validity, should be addressed given the wide-spread implementation of the tool in routine clinical practice. Furthermore, the moderate level of consistency between different informants indicate that an assessment of a pre-schooler should not rely on a single informant. Further work is required to examine these psychometric properties in parallel with qualitative work that can explore acceptability and validity of the SDQ in more depth. Whilst such evidence is gathered it remains critical to not solely rely on SDQ scores but also consider parents' and teachers' reports before determining the needs of pre-school children.

Funding

The review was funded by the <Name and web address of Centre>

Conflict of interest

The authors do not have a conflict of interest

Box 1. Definitions used of measurement properties (Mokkink et al., 2010) (p261-3 (Streiner & Norman, 2008) (R. Goodman, 1997)

Content validity:

- The degree to which the content of an instrument is an adequate reflection of the construct to be measured

Construct validity:

- Discriminative validity: ability of a tool to discriminate between two extreme groups
- Convergent validity: The degree to which the scores of the (new) scale relate to scores on other measures to which it should be related
- Discriminant/divergent validity: The degree to which the scores of the (new) scale do not relate to scores on another scale that measures dissimilar constructs
- Structural validity: The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured
- Cross-cultural validity: The degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the

original version of the instrument

Criterion validity:

- Concurrent validity: The correlation of the instrument with a "gold standard" criterion administered at the same time
- Predictive validity: The correlation of the instrument with a "gold standard" criterion that will be available in the future

Internal consistency:

- The degree of the interrelatedness among the items

Reliability

- Intra-rater reliability: The extent to which scores for people who have not changed are the same for repeated measurement by the same rater
- Inter-rater reliability: The extent to which scores for people who have not changed are the same for repeated measurement by different raters (of the same type) on the same occasion
- Cross-informant consistency: The extent to which scores for people who have not changed are the same for repeated measurement by different types of raters on the same occasion
- Test-retest reliability: The extent to which scores on the same version of questionnaire for people who have not changed are the same for repeated measurement over time

Measurement error:

- The systematic and random error of a person's score that is not attributed to true changes in the construct to be measured

Responsiveness:

- The ability of an instrument to detect change over time in the construct to be measured

Figure 1 Literature search results

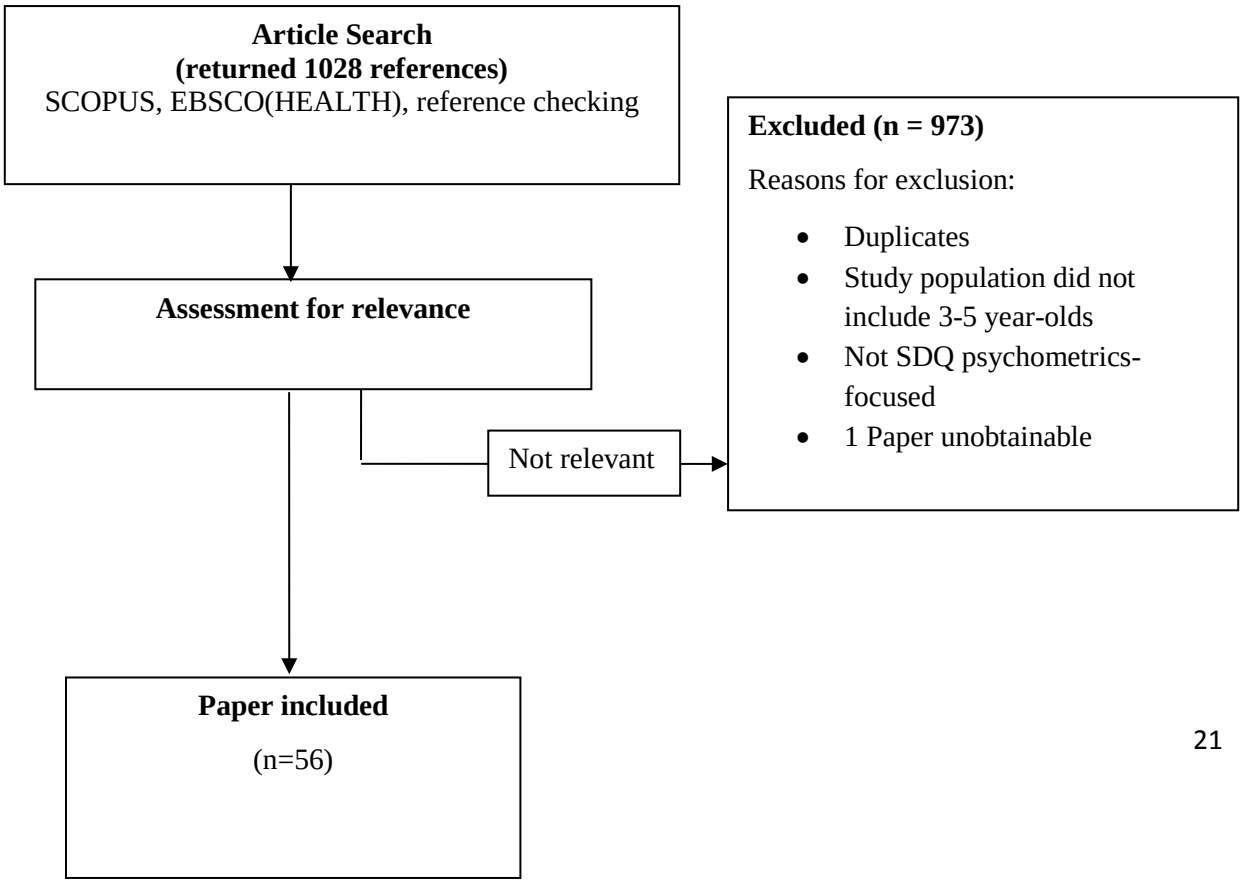

Table 1. Area Under the Curve (AUC) values from studies examining discriminative validity

Source	Study quality *	Age group (years)	Sample size (n)	Emotional Symptoms	Conduct Problems	Hyper-activity	Peer Problems	Prosocial Behaviour	Total Difficulty	Comparison groups
Parents' questionnaires										
(Alyahri & Goodman, 2006)	Fair	5-12	197	0.78	0.88	0.97	0.7	0.78	0.81	Clinic vs community
(Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004)	Fair	5-17	543	0.69	0.81	0.76	__	__	0.77	Clinic vs community
(Du et al., 2008)	Fair	3-17	94	__	0.68	0.77	0.55	0.39	0.69	Community vs Attention Deficit / Hyperactivity Disorder cases
(R. Goodman, 1997)	Fair	4-16	403	__	__	__	__	__	0.87	Clinic vs community
(R. Goodman & Scott, 1999)	Fair	4-7	132	0.71	0.92	0.86	0.75	__	0.93	Clinic vs community
(Klasen et al., 2000)	Fair	4-16	273	0.85	0.97	0.94	0.78	__	0.91	Clinic vs community
(Malmberg, 2003)	Fair	5-15	493	0.81	0.81	0.81	0.76	0.72	0.89	Clinic vs community
(Samad et al., 2005)	Fair	4-16	212	0.74	__	0.8	0.73	0.82	0.77	Psychiatric vs paediatric clinics
(Sveen, Berg-Nielsen, Lydersen, & Wichstrøm, 2013)	Fair	4	845	__	__	__	__	__	0.75	Cases vs non-cases based on a psychiatric interview, any disorder
Teachers' questionnaires										
(Alyahri & Goodman, 2006)	Fair	5-12	197	0.7	0.86	0.97	0.66	0.65	0.76	Clinic vs community
(Becker et al., 2004)	Fair	5-17	543	0.65	0.82	0.79	__	__	0.75	Clinic vs community
(Du et al., 2008)	Fair	3-17	94	__	0.87	0.9	0.69	0.67	0.91	Community vs Attention Deficit / Hyperactivity Disorder cases
(R. Goodman, 1997)	Fair	4-16	403	__	__	__	__	__	0.85	Clinic vs community

(Sveen et al., 2013)	Fair	4	845	__	__	__	__	__	0.64	Cases vs non-cases based on a psychiatric interview, any disorder

* Assessed with the COSMIN tool

Table 2. Summary findings from studies examining convergent validity

Strength and Difficulties Questionnaire version	Number of studies (analyses) *	Median sample size (range)	Emotional Symptoms	Conduct Problems	Hyper-activity	Peer Problems	Prosocial Behaviour	Total Difficulty
Parents' questionnaires	14 (19)	292 (29-1940)						
Weighted Average (all comparators; indicative only)			0.55	0.58	0.63	0.5	-0.18	0.67
Between-study Weighted Standard Deviation			(0.15)	(0.18)	(0.09)	(0.14)	(0.04)	(0.12)
Teachers' questionnaires	8 (11)	179 (21-543)						
Weighted Average (all comparators; indicative only)			0.66	0.78	0.8	0.54	-0.35	0.78
Between-study Weighted Standard Deviation			(0.16)	(0.08)	(0.05)	(0.12)	(0.16)	(0.11)

* Full data extracted are provided in Supplementary file A (Becker et al., 2004; Birkás et al., 2008; Downs, Strand, Heinrichs, & Cerna, 2012; Du et al., 2008; Ezpeleta et al., 2012; R. Goodman, 1997; R. Goodman & Scott, 1999; Hawes & Dadds, 2004; Janssens, 2009; Klasen et al., 2000; Mathai, Anderson, & Bourne, 2002; C. Mieloo et al., 2012; C. L. Mieloo et al., 2014; Petermann et al., 2010; Theunissen, Vogels, De Wolff, & Reijneveld, 2013; Van Leeuwen, Meerschaert, Bosmans, De Medts, & Braet, 2006)

Table 3. Summary findings from studies examining structural validity

Source	Study quality*	Analytical approach used **	Age group (years)	Note	Sample size (n)	Root Mean Square Approximation	Confirmatory Fit Index
Parents' questionnaires							
(Williamson et al., 2014)	Fair	CFA (SEM)	4-17		717	0.07	0.81
(Gómez-Beneyto et al., 2013)	Excellent	CFA	4-15		3,253	0.06	0.96
(Becker et al., 2004)	Excellent	CFA	5-17		543	—	—
(Dickey & Blumberg, 2004)	Excellent	CFA (polychoric correlations)	4-17		4,804	—	—
(Ezpeleta et al., 2012)	Fair	CFA	3		1,341	0.03	0.88
(A. Goodman, Lamping, et al., 2010)	Excellent	CFA	5-16		18,222	0.06	0.86
(A. Goodman, Patel, et al., 2010)	Excellent	CFA	5-16		14,229	0.05	0.96
(Hill & Hughes, 2007)	Fair	CFA	5-7		505	0.07	0.82
(Hill & Hughes, 2007)	Fair	SEM	5-7		505	0.05	0.87
(Klein, Otto, Fuchs, Zenger, & Von Klitzing, 2013)	Fair	CFA	3-5		1,738	0.05	0.86
(C. Mieloo et al., 2012)	Excellent	CFA	5-6		4,325	0.05	0.88
(Niclasen, Skovgaard, et al., 2012)	Excellent	CFA SEM	5-7	Boys	28,920	0.03	0.89
(Niclasen, Skovgaard, et al., 2012)	Excellent	CFA SEM	5-7	Girls	27,611	0.03	0.91
(Theunissen et al., 2013)	Fair	SEM–based CFA	3-4		839	0.08	—
(Van Leeuwen et al., 2006)	Excellent	CFA	4-8	Sample 1	532	0.06	0.89
(Van Leeuwen et al., 2006)	Excellent	CFA	4-8	Sample 2	1,086	0.08	0.89
Custodial grandparents' questionnaires							
(Palmieri & Smith, 2007)	Fair	SEM–based CFA	4-16		733	0.06	0.95
Teachers' questionnaires							
(Tobia, Gabriele, & Marzocchi, 2013)	Fair	CFA	3-15		2,302	0.07	0.94
(Downs et al., 2012)	Fair	CFA tetrachoric correlation	3-5	Spanish	129	0.05	0.92

(Downs et al., 2012)	Fair	CFA tetrachoric correlation	3-5	English	169	0.07	0.91
(Downs et al., 2012)	Fair	CFA tetrachoric correlation	3-6	German	179	0.05	0.90
(Ezpeleta et al., 2012)	Fair	CFA	3		622	0.03	0.88
(A. Goodman, Lamping, et al., 2010)	Excellent	CFA	5-16		14,263	0.09	0.91
(A. Goodman, Patel, et al., 2010)	Excellent	CFA	5-16		11,032	0.06	0.99
(Hill & Hughes, 2007)	Fair	CFA	5-7		676	0.08	0.87
(Hill & Hughes, 2007)	Fair	SEM	5-7		676	0.07	0.89
(C. Mieloo et al., 2012)	Excellent	CFA	5-6		4,314	0.07	0.89
(Niclasen, Skovgaard, et al., 2012)	Excellent	CFA (SEM)	5-7	Boys	1,272	0.05	0.96
(Niclasen, Skovgaard, et al., 2012)	Excellent	CFA (SEM)	5-7	Girls	1,291	0.05	0.96
(Van Leeuwen et al., 2006)	Excellent	CFA	4-8	Sample1	512	0.06	0.9
(Van Leeuwen et al., 2006)	Excellent	CFA	4-8	Sample 2	1,049	0.08	0.92

* Assessed with the COSMIN tool

** All fit statistics reported in the studies are provided in supplementary file B. CFA = Confirmatory Factor Analysis; SEM = Structural Equation Modelling

Table 4. Summary findings from studies examining internal consistency (Chronbach's Alpha)

Strength and Difficulties Questionnaire version	Number of studies (analyses) *	Median sample size (range)	Emotional Symptoms	Conduct Problems	Hyper-activity	Peer Problems	Prosocial Behaviour	Total Difficulty
Parents' questionnaires	22 (29)	733 (156-22,018)						
Weighted Average			**0.62**	**0.56**	**0.69**	**0.49**	**0.66**	**0.76**
Between-study Weighted S.D.			**0.06**	**0.06**	**0.07**	**0.20**	**0.03**	**0.07**
Teachers' questionnaires	16 (22)	761 (129-14,263)						
Weighted Average			**0.75**	**0.70**	**0.82**	**0.69**	**0.83**	**0.82**
Between-study Weighted S.D.			**0.04**	**0.07**	**0.05**	**0.14**	**0.01**	**0.07**

* Data from all included studies are provided in Supplementary file C (Becker et al., 2004; Birkás et al., 2008; Borg et al., 2012; Bourdon et al., 2005; Cefaia et al., 2011; Downs et al., 2012; Du et al., 2008; Gao, Shi, Zhai, He, & Shi, 2013; A. Goodman, Lamping, et al., 2010; R. Goodman, 2001; Hawes & Dadds, 2004; Hayes, 2007; Hill & Hughes, 2007; Janssens, 2009; Klein et al., 2013; Koglin, Barquero, Mayer, Scheithauer, & Petermann, 2007; Malmberg, 2003; Matsuishi et al., 2008; C. Mieloo et al., 2012; C. L. Mieloo et al., 2014; Palmieri & Smith, 2007; Sveen et al., 2013; Theunissen et al., 2013; Tobia et al., 2013; Van Leeuwen et al., 2006; Williamson et al., 2014).

Table 5. Summary findings from studies examining cross-informant consistency

Source	Study quality*	Statistic used	Age group (years)	Respondents	Sample size n	Emotional Symptoms	Conduct Problems	Hyper-activity	Peer Problems	Prosocial Behaviour	Total Difficulty
(Downs et al., 2012)	Fair	ICC	3-6	Father, mother & teacher	111	0.36	0.33	0.5	0.39	0.25	0.43
(Du et al., 2008)	Good	Pearson ρ	3-17	Parent & teacher	1,965	0.23	0.31	0.44	0.29	0.27	0.36
(R. Goodman, 1997)	Fair	Pearson ρ	4-16	Parent & teacher	128	0.41	0.65	0.54	0.59	0.37	0.62
(R. Goodman, 2001)	Fair	Pearson ρ	5-15	Parent & teacher	7,313	0.27	0.37	0.48	0.37	0.25	0.46
(Janssens, 2009)	Fair	Pearson ρ	3-18	Parent & caregiver	206	0.43	0.6	0.58	0.45	0.35	0.57
(Mathai et al., 2002)	Fair	Pearson ρ	4-14	Parent & teacher	99	0.35	0.41	0.61	0.45	__	0.37
(C. Mieloo et al., 2012)	Fair	Pearson ρ	5-6	Parent & teacher	3,718	0.29	0.25	0.45	0.29	0.22	0.41
(C. L. Mieloo et al., 2014)	Fair	Pearson ρ	5-6	Parent & teacher, Surinamese	435	0.11	0.22	0.41	0.23	0.12	0.28
(C. L. Mieloo et al., 2014)	Fair	Pearson ρ	5-6	Parent & teacher, Antillean/Aruban	207	0.11	0.27	0.4	0.22	0.32	0.32
(C. L. Mieloo et al., 2014)	Fair	Pearson ρ	5-6	Parent & teacher, Turkish	535	0.13	0.17	0.31	0.18	0.18	0.23
(C. L. Mieloo et al., 2014)	Fair	Pearson ρ	5-6	Parent & teacher, Moroccan	516	0.09	0.16	0.29	0.08	0.12	0.2
(Niclasen, Skovgaard, et al., 2012)	Fair	Pearson ρ	5	Parent & teacher	2,594	0.33	0.33	0.42	0.37	0.29	0.45
(Thabet et al., 2000)	Fair	Spearman ρ	3-16	Parent & teacher	322	0.21	0.21	0.18	__	__	0.2
(Van Leeuwen et al., 2006)	Fair	Pearson ρ	4-7	Parent & teacher, Sample 1	512	0.26	0.31	0.5	0.27	0.22	__
(Van Leeuwen et al., 2006)	Fair	Pearson ρ	4-7	Parent & teacher, Sample 2	1,049	0.29	0.27	0.45	0.25	0.27	__
			Total sample size for property			**19710**	**19710**	**19710**	**19388**	**19289**	**18149**
			Weighted average correlation			**0.27**	**0.32**	**0.45**	**0.32**	**0.25**	**0.41**
			Between-study weighted SD			**0.06**	**0.07**	**0.06**	**0.07**	**0.04**	**0.07**

* Assessed with the COSMIN tool

Table 6. Summary findings from studies examining test-retest reliability

Source	Study quality*	Statistic used	Age group (years)	Participants	Sample size n	Emotional Symptoms	Conduct Problems	Hyper-activity	Peer Problems	Prosocial Behaviour	Total Difficulty
(Borg et al., 2012)	Fair	Spearman's ρ	4-9	Parent	592	**	**	0.79	**	__	0.76
(Downs et al., 2012)	Fair	Pearson's partial correlation	3-5	Teacher rating child with English as first language	264	0.62	0.5	0.67	0.55	0.56	0.61
(Downs et al., 2012)	Fair	Pearson's partial correlation	3-5	Teacher rating child with Spanish as first language	264	0.38	0.52	0.71	0.53	0.6	0.66
(Du et al., 2008)	Fair	Correlations	3-17	Parent	45	0.47	0.7	0.48	0.79	0.43	0.72
(Du et al., 2008)	Fair	Correlations	3-17	Teacher	45	0.4	0.5	0.64	0.58	0.5	0.55
(R. Goodman, 1999)	Fair	ICC	5-15	Parent	34	__	__	__	__	__	0.85
(R. Goodman, 2001)	Fair	Correlations	5-15	Parent	2091	0.57	0.64	0.72	0.61	0.61	0.72
(R. Goodman, 2001)	Fair	Correlations	5-15	Teacher	796	0.65	0.69	0.82	0.72	0.74	0.8
(Hawes & Dadds, 2004)	Poor	Pearson's correlations	4-9	Parent	780	0.71	0.65	0.77	0.61	0.64	0.77

* Assessed with the COSMIN tool
** The authors presented a range of correlations for these 3 subscales (.6 to.68)

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin, 101*(2), 213-232. doi: 10.1037/0033-2909.101.2.213

Alyahri, A., & Goodman, R. (2006). Validation of the Arabic Strengths and Difficulties Questionnaire and the Development and Well-Being Assessment. *Eastern Mediterranean Health Journal = La Revue De Santé De La Méditerranée Orientale = Al-Majallah Al-Ṣiḥḥīyah Li-Sharq Al-Mutawassiṭ, 12 Suppl 2*, S138-S146.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3186-3191.

Becker, A., Woerner, W., Hasselhorn, M., Banaschewski, T., & Rothenberger, A. (2004). Validation of the parent and teacher SDQ in a clinical sample. *European Child & Adolescent Psychiatry, 13*, ii11-ii16. doi: 10.1007/s00787-004-2003-5

Bekker, J., Bruck, D., & Sciberras, E. (2013). Congruent Validity of the Strength and Difficulties Questionnaire to Screen for Comorbidities in Children With ADHD. *Journal Of Attention Disorders*.

Bierman, K. L., Coie, J., Dodge, K., Greenberg, M., Lochman, J., McMohan, R., . . . McMahon, R. J. (2013). School Outcomes of Aggressive-Disruptive Children: Prediction From Kindergarten Risk Factors and Impact of the Fast Track Prevention Program. *Aggressive Behavior, 39*(2), 114-130.

Birkás, E., Lakatos, K., Tóth, I., & Gervai, J. (2008). [Screening childhood behavior problems using short questionnaires I.: the Hungarian version of the Strengths and Difficulties Questionnaire]. *Psychiatria Hungarica : A Magyar Pszichiatriai Tarsasag tudomanyos folyoirata, 23*(5), 358-365.

Borg, A.-M., Pälvi, K., Raili, S., Matti, J., & Tuula, T. (2012). Reliability of the Strengths and Difficulties Questionnaire among Finnish 4-9-year-old children. *Nordic Journal of Psychiatry, 66*(6), 403-413. doi: 10.3109/08039488.2012.660706

Bourdon, K. H., Goodman, R., Rae, D. S., Simpson, G., & Koretz, D. S. (2005). The Strengths and Difficulties Questionnaire: U.S. normative data and psychometric properties. *Journal Of The American Academy Of Child And Adolescent Psychiatry, 44*(6), 557-564.

Caspi, A., Moffitt, T. E., Newman, D. L., & Silva, P. A. (1996). Behavioral observations at age 3 years predict adult psychiatric disorders: Longitudinal evidence from a birth cohort. *Archives of General Psychiatry, 53*(11), 1033-1039.

Cefaia, C., Camillerib, L., Cooperc, P., & Saidd, L. (2011). The structure and use of the teacher and parent Maltese Strengths and Difficulties Questionnaire.

Critical Appraisal Skills Programme (CASP). (2003). *Critical Appraisal Skills Programme: appraisal tools*. Public Health Resource Unit.

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*: Pearson Education.

De Giacomo, A., Lamanna, A. L., Craig, F., Santoro, N., Goffredo, S., & Cecinati, V. (2012). The SDQ in Italian clinical practice: evaluation between three outpatient groups compared. *Rivista di psichiatria, 47*(5), 400-406.

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the strengths and difficulties questionnaire: United States, 2001. *Journal Of The American Academy Of Child And Adolescent Psychiatry, 43*(9), 1159-1167.

Doughty, C. (2005). The effectiveness of mental health promotion, prevention and early intervention in children, adolescents and adults. NZHTA Report ; 8(2).

Downs, A., Strand, P. S., Heinrichs, N., & Cerna, S. (2012). Use of the Teacher Version of the Strengths and Difficulties Questionnaire with German and American Preschoolers. *Early Education and Development, 23*(4), 493-516.

Du, Y., Kou, J., & Coghill, D. (2008). The validity, reliability and normative scores of the parent, teacher and self report versions of the Strengths and Difficulties Questionnaire in China. *Child And Adolescent Psychiatry And Mental Health, 2*.

Eivers, A. R., Brendgen, M., & Borge, A. I. H. (2010). Stability and change in prosocial and antisocial behavior across the transition to school: Teacher and peer perspectives. *Early Education and Development, 21*(6), 843-864.

Ezpeleta, L., Granero, R., la Osa, N. d., Penelo, E., & Domènech, J. M. (2012). Psychometric properties of the Strengths and Difficulties Questionnaire(3-4) in 3-year-old preschoolers. *Comprehensive Psychiatry*.

Gao, X., Shi, W., Zhai, Y., He, L., & Shi, X. (2013). Results of the parent-rated Strengths and Difficulties Questionnaire in 22, 108 primary school students from 8 provinces of China. *Shanghai Archives of Psychiatry, 25*(6), 364-374.

Gómez-Beneyto, M., Nolasco, A., Moncho, J., Pereyra-Zamora, P., Tamayo-Fonseca, N., Munarriz, M., . . . Girón, M. (2013). Psychometric behaviour of the strengths and difficulties questionnaire (SDQ) in the Spanish national health survey 2006. *BMC Psychiatry, 13*.

Goodman, A., & Goodman, R. (2009). Strengths and difficulties questionnaire as a dimensional measure of child mental health. *Journal Of The American Academy Of Child And Adolescent Psychiatry, 48*(4), 400-403.

Goodman, A., & Goodman, R. (2011). Population mean scores predict child mental disorder rates: Validating SDQ prevalence estimators in Britain. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 52*(1), 100-108.

Goodman, A., Heiervang, E., Fleitlich-Bilyk, B., Alyahri, A., Patel, V., Mullick, M. S. I., . . . Goodman, R. (2012). Cross-national differences in questionnaires do not necessarily reflect comparable differences in disorder prevalence. *Social Psychiatry And Psychiatric Epidemiology, 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. *Journal Of Abnormal Child Psychology, 38*(8), 1179-1191.

Goodman, A., Patel, V., & Leon, D. A. (2010). Why do British Indian children have an apparent mental health advantage? *Journal of Child Psychology and Psychiatry and Allied Disciplines, 51*(10), 1171-1183.

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 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. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 40*(5), 791-799.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal Of The American Academy Of Child And Adolescent Psychiatry, 40*(11), 1337-1345.

Goodman, R., Ford, T., Corbin, T., & Meltzer, H. (2004). Using the Strengths and Difficulties Questionnaire (SDQ) multi-informant algorithm to screen looked-after children for psychiatric disorders. *European Child and Adolescent Psychiatry, Supplement, 13*(2), II/25-II/31.

Goodman, R., Meltzer, H., & Bailey, V. (1998). The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry, 7*(3), 125-130.

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. *European Child and Adolescent Psychiatry, 9*(2), 129-134.

Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the child behavior checklist: Is small beautiful? *Journal Of Abnormal Child Psychology, 27*(1), 17-24.

Gruenert, S., Ratnam, S., & Tsantefski, M. (2006). Identifying children's needs when parents access drug treatment: The utility of a brief screening measure. *Journal of Social Work Practice in the Addictions, 6*(1-2), 139-154.

Hawes, D. J., & Dadds, M. R. (2004). Australian data and psychometric properties of the Strengths and Difficulties Questionnaire. *Australian and New Zealand Journal of Psychiatry, 38*(8), 644-651.

Hayes, L. (2007). Problem behaviours in early primary school children: Australian normative data using the Strengths and Difficulties Questionnaire. *Australian & New Zealand Journal of Psychiatry, 41*(3), 231-238.

Hill, C. R., & Hughes, J. N. (2007). An Examination of the Convergent and Discriminant Validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly: The Official Journal Of The Division Of School Psychology, American Psychological Association, 22*(3), 380-406.

Holtmann, M., Becker, A., Banaschewski, T., Rothenberger, A., & Roessner, V. (2011). Psychometric validity of the strengths and difficulties questionnaire-dysregulation profile. *Psychopathology, 44*(1), 53-59.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Janssens, A. D. (2009). Screening for psychopathology in child welfare: the Strengths and Difficulties Questionnaire (SDQ) compared with the Achenbach System of Empirically Based Assessment (ASEBA). *European Child & Adolescent Psychiatry, 18*(11), 691-700. doi: 10.1007/s00787-009-0030-y

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry, 62*(6), 593-602.

Kim-Cohen, J., Arseneault, L., Newcombe, R., Adams, F., Bolton, H., Cant, L., . . . Moffitt, T. E. (2009). Five-year predictive validity of DSM-IV conduct disorder research diagnosis in 41/2-5-year-old children. *European Child and Adolescent Psychiatry, 18*(5), 284-291.

Kim-Cohen, J., Caspi, A., Moffitt, T. E., Harrington, H., Milne, B. J., & Poulton, R. (2003). Prior juvenile diagnoses in adults with mental disorder: Developmental follow-back of a prospective-longitudinal cohort. *Archives of General Psychiatry, 60*(7), 709-717.

Klasen, H., Woerner, W., Wolke, D., Meyer, R., Overmeyer, S., Kaschnitz, W., . . . Goodman, R. (2000). Comparing the German versions of the Strengths and Difficulties Questionnaire (SDQ-Deu) and the Child Behavior Checklist. *European Child and Adolescent Psychiatry, 9*(4), 271-276.

Klein, A. M., Otto, Y., Fuchs, S., Zenger, M., & Von Klitzing, K. (2013). Psychometric properties of the parent-rated SDQ in preschoolers. *European Journal of Psychological Assessment, 29*(2), 96-104.

Koglin, U., Barquero, B., Mayer, H., Scheithauer, H., & Petermann, F. (2007). German version of the Strengths and Difficulties Questionnaire (T4-16 - SDQ): Psychometric quality of the teacher version for preschoolers. *Diagnostica, 53*(4), 175-183.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of Internal Medicine, 151*(4), W65-94.

Malmberg, M. A.-M. H. (2003). Validity of the Swedish version of the Strengths and Difficulties Questionnaire (SDQ-Swe). *Nordic Journal of Psychiatry, 57*(5), 357. doi: 10.1080/08039480310002697

Mathai, J., Anderson, P., & Bourne, A. (2002). The Strengths and Difficulties Questionnaire (SDQ) as a screening measure prior to admission to a Child and Adolescent Mental Health Service (CAMHS). *Advances in Mental Health, 1*(3), 235-246.

Mathai, J., Anderson, P., & Bourne, A. (2003). Use of the Strengths and Difficulties Questionnaire as an outcome measure in a child and adolescent mental health service. *Australasian Psychiatry, 11*(3), 334-337.

Mathai, J., Anderson, P., & Bourne, A. (2004). Comparing psychiatric diagnoses generated by the Strengths and Difficulties Questionnaire with diagnoses made by clinicians. *Australian and New Zealand Journal of Psychiatry, 38*(8), 639-643.

Matsuishi, T., Nagano, M., Araki, Y., Tanaka, Y., Iwasaki, M., Yamashita, Y., . . . Kakuma, T. (2008). Scale properties of the Japanese version of the Strengths and Difficulties Questionnaire (SDQ): A study of infant and school children in community samples. *Brain and Development, 30*(6), 410-415.

Mieloo, C., Raat, H., van Oort, F., Bevaart, F., Vogel, I., Donker, M., & Jansen, W. (2012). Validity and reliability of the strengths and Difficulties Questionnaire in 5-6 year olds: Differences by gender or by parental education? *Plos One, 7*(5).

Mieloo, C. L., Bevaart, F., Donker, M. C., van Oort, F. V., Raat, H., & Jansen, W. (2014). Validation of the SDQ in a multi-ethnic population of young children. *The European Journal of Public Health, 24*(1), 26-32. doi: 10.1093/eurpub/ckt100

Ministry of Health. (2008). The B4 School Check. A handbook for practitioners. Wellington.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737-745.

Niclasen, J., Skovgaard, A. M., Andersen, A.-M. N., Sømhovd, M. J., & Obel, C. (2012). A Confirmatory Approach to Examining the Factor Structure of the Strengths and Difficulties Questionnaire (SDQ): A Large Scale Cohort Study. *Journal Of Abnormal Child Psychology*.

Niclasen, J., Teasdale, T. W., Andersen, A.-M. N., Skovgaard, A. M., Elberling, H., & Obel, C. (2012). Psychometric properties of the Danish Strength and Difficulties Questionnaire: the SDQ assessed for more than 70,000 raters in four different cohorts. *Plos One, 7*(2), e32025-e32025. doi: 10.1371/journal.pone.0032025

Palmieri, P. A., & Smith, G. C. (2007). Examining the Structural Validity of the Strengths and Difficulties Questionnaire (SDQ) in a U.S. Sample of Custodial Grandmothers. *Psychological Assessment, 19*(2), 189-198.

Petermann, U., Petermann, F., & 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.

Samad, L., Hollis, C., Prince, M., & Goodman, R. (2005). Child and adolescent psychopathology in a developing country: Testing the validity of the Strengths and Difficulties Questionnaire (Urdu version). *International Journal Of Methods In Psychiatric Research, 14*(3), 158-166.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-year-olds: a review. *Clinical Child And Family Psychology Review, 13*(3), 254-274. doi: 10.1007/s10567-010-0071-2

Streiner, D. L., & Norman, G. R. (2008). *Health Measurement Scales: a practical guide to their development and use*. Oxford: Oxford University Press.

Stringaris, A., & Goodman, R. (2013). The value of measuring impact alongside symptoms in children and adolescents: A longitudinal assessment in a community sample. *Journal Of Abnormal Child Psychology, 41*(7), 1109-1120.

Sveen, T. H., Berg-Nielsen, T. S., Lydersen, S., & Wichstrøm, L. (2013). Detecting psychiatric disorders in preschoolers: Screening with the strengths and difficulties questionnaire. *Journal Of The American Academy Of Child And Adolescent Psychiatry, 52*(7), 728-736.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., . . . de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34-42.

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research, 21*(4), 651-657. doi: 10.1007/s11136-011-9960-1

Thabet, A. A., Stretch, D., & Vostanis, P. (2000). Child mental health problems in Arab children: Application of the strengths and difficulties questionnaire. *International Journal of Social Psychiatry, 46*(4), 266-280.

Theunissen, M. H. C., Vogels, A. G. C., De Wolff, M. S., & Reijneveld, S. A. (2013). Characteristics of the strengths and difficulties questionnaire in preschool children. *Pediatrics, 131*(2), e446-e454.

Thomson Reuters. (2010). EndNote X4.

Tobia, V., Gabriele, M. A., & Marzocchi, G. M. (2013). The Italian Version of the Strengths and Difficulties Questionnaire (SDQ)-Teacher: Psychometric Properties. *Journal of Psychoeducational Assessment, 31*(5), 493-505.

Toh, T. H., Chow, S. J., Ting, T. H., & Sewell, J. (2008). Chinese translation of strengths and difficulties questionnaire requires urgent review before field trials for validity and reliability. *Child And Adolescent Psychiatry And Mental Health, 2*, 23.

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.

White, J., Connelly, G., Thompson, L., & Wilson, P. (2013). Assessing wellbeing at school entry using the Strengths and Difficulties Questionnaire: Professional perspectives. *Educational Research, 55*(1), 87-98.

Williamson, A., McElduff, P., Dadds, M., D' Este, C., Redman, S., Raphael, B., . . . Eades, S. (2014). The Construct Validity of the Strengths and Difficulties Questionnaire for Aboriginal Children Living in Urban New South Wales, Australia. *Australian Psychologist*.

Williamson, A., Redman, S., Dadds, M., Daniels, J., D'Este, C., Raphael, B., . . . Skinner, T. (2010). Acceptability of an emotional and behavioural screening tool for children in Aboriginal Community Controlled Health Services in urban NSW. *Australian and New Zealand Journal of Psychiatry, 44*(10), 894-900.