The Ability of the Strength and Difficulties Questionnaire to Detect Mental Health Disorders in a Child and Adolescent Outpatient Clinic in Northern-Norway

Therese Fjeldmo Moe og Guri Sæther

Veiledere:
Martin Eisemann
Per Håkan Brøndbo
Børge Mathiassen

PSY 2901

Hovedoppgave for graden Cand.Psychol.
Fakultet for helsevitenskap, Institutt for Psykologi
Universitetet i Tromsø
Vår 2010

Forord

Datamaterialet til denne studien fikk vi fra barne- og ungdomspsykiatrisk poliklinikk (BUP) i Tromsø. Dette materialet er en del av en større regional undersøkelse, BUP-Nord studien. Vi er takknemlige ovenfor våre biveiledere tilknyttet BUP Tromsø for å ha gjort datamaterialet tilgjengelig for oss.

Årsaken for valg av tema er vår store interesse for barne- og ungdomspsykiatri. Ideen til problemstilling fikk vi fra biveilederne våre ved BUP Tromsø, og denne ble videreutviklet av oss. Vi har jobbet sammen under hele prosessen, fra utforming av problemstilling til ferdigstilt oppgave. Det meste av relevant litteratur til problemstillingen har vi selv funnet, med noen tips fra våre veiledere. De statistiske analysene og det skriftlige arbeidet har vi utført selv.

Vi vil takke vår hovedveileder Martin Eisemann ved Institutt for Psykologi for å ha korrekturlest oppgaven, og kommet med positive tilbakemeldinger og støtte underveis. Vi vil også rette en takk til våre biveiledere Børge Mathiassen og Per Håkan Brøndbo ved BUP Tromsø for interessante og konstruktive diskusjoner, samt tilbakemeldinger på innholdet i oppgaven.

Gjennom prosessen med hovedoppgaveskriving har vi lært veldig mye, og er blitt inspirert til å jobbe videre innen barne- og ungdomspsykiatrien.

Therese Fjeldmo Moe og Guri Sæther

Tromsø, 03.05.2010

Abstract

The aim of the study was to examine the ability of the Strength and Difficulties Questionnaire (SDQ) to detect mental health disorders among patients referred to a child and adolescent outpatient clinic. The sample consisted of 100 participants between 5 and 17 years referred to an outpatient clinic in Northern-Norway. The SDQ-generated diagnostic predictions were compared to a gold standard to measure the screening efficiency of the SDQ. The gold standard consisted of psychiatric diagnoses assigned by experienced clinicians on the basis of the Developmental and Well-Being Assessment (DAWBA). Screening efficiency was measured by using sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios. The results showed that 72% of the children and adolescents in the sample had a diagnosable mental disorder according to the SDQ. In terms of screening efficiency the findings showed mixed results for the different diagnostic disorders. Overall the results suggest that the SDQ might not be accurate enough to be of practical use as a screening instrument in a child and adolescent mental health clinic.

Mental health problems are very common in children and adolescents. Approximately 15 to 20% of Norwegian children and adolescents suffer from psychosocial impairment due to such problems (Mathiesen, 2009). Children and adolescents with mental health problems in Norway are often referred to child and adolescent outpatient clinics. At these outpatient clinics the patients are evaluated on the basis of clinical judgments at intake, to assess the severity of their difficulties according to definitions in Norwegian legislation (Helsedirektoratet, 2008). This evaluation is aimed to distinguish between those who have a mental health problem and those who actually have a mental health disorder. The outpatient clinics are obligated to offer necessary counselling and treatment to patients with disorders as well as to their families (Andersson, Ose, & Sitter, 2005).

The Norwegian Institute of Public Health (Mathiesen, 2009) estimates a prevalence of psychiatric disorders among 3-to-18-year-olds of about 8%. On the other hand, the percentage of children and adolescent receiving mental health care is generally lower. The coverage of services given is only 4.5% on a national basis (Mathiesen, 2009). The gap between the percentage of children in need of mental health services and the actual coverage highlights the capacity problem in Norwegian child and adolescent mental health care. Long waiting lists are one example of capacity problems making it more difficult for children and adolescents to get the appropriate help needed (Andersson et al., 2005). Similar problems are experienced internationally (Mathai, Anderson, & Bourne, 2004). Increased accessibility for the services has been called for (Andersson, 2009), and to achieve this there has been a focus on increased productivity in the Norwegian clinics. One way to attain increased productivity is to make the intake process more efficient. Efficiency in the intake process can include both a reduction of time spent, as well as trying to avoid the intake of patients without a mental health disorder. If this can be done more children and adolescents in need of mental health service might get appropriate help. To improve the efficiency in the intake process, more rapid first-assessments regarding the type and severity of disorders might be used. Using standardised methods such as screening instruments is one way of identifying whether a disorder is present or not and if further evaluation is required (Warner, 2004). Ægisdóttir and colleagues (2006) compared the effect of standardised methods to clinical judgment alone, and reported that somewhat greater accuracy was found for standardised methods. A

guide for the child and adolescent mental health services (Helsedirektoratet, 2008) also recommends the use of standardised methods in addition to clinical judgment.

Many different screening instruments are available, and the use of these instruments varies across different services and regions (Helsedirektoratet, 2008). One of these screening instruments is the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997; Goodman, 1999). The SDQ is a screening device for assessing the behaviour of children and adolescents, based on information from parents, teachers and self-report. It is widely used to screen for psychiatric disorders, as a measure of treatment outcome in mental health clinics, and as a research instrument (Garralda, Yates, & Higginson, 2000; Goodman, Ford, Simmons, Gatward, & Meltzer, 2000b; Goodman, Renfrew, & Mullick, 2000c). Several studies, both internationally (Bourdon, Goodman, Donald, Simpson, & Koretz, 2005; Mellor, 2004; Woerner et al., 2004) and in the Nordic countries (Rønning, Handegaard, Sourander, & Mørch, 2004; Smedje, Broman, Hetta, & von-Knorring, 1999), have aimed towards establishing norms and evaluating the psychometric properties of the SDQ. In a review by Obel and colleagues (2006), the authors report that although several studies have been conducted in Norway to examine the psychometric properties of the SDQ, these studies have not used all three informants or covered the entire age range and are merely based on community samples.

There are several advantages by using the SDQ. Firstly, it focuses on the children's and adolescents' strengths as well as difficulties (Goodman, 1997). Symptoms of a disorder do not always reflect the strain a person is experiencing, thus the SDQ also includes an impact supplement in addition to the symptom score (Goodman, 1999). The impact supplement covers perceived severity of the problems, including overall distress, social impairment, burden, and chronicity. Another important advantage of the SDQ is the use of multiple sources to assess children's and adolescents' mental health (Heyerdahl, 2003). For example, some behavioural problems can be highly situational, and to meet the criteria for psychiatric diagnoses such as ADHD/hyperkinesis, symptoms and impairment related to the problem are required to be reported in at least two settings (American Psychiatric Association, 1994; World Health Organization, 1996). In addition, the adolescents' self-report is important because their perception of their own problems may be different from their parents' and teachers' (Heyerdahl, 2003). When

examining the inter-informant reliability, Mellor (2004) and Mathai, Anderson and Bourne (2002) reported moderate inter-informant agreement. A Norwegian study reported moderate to high inter-informant agreement between parents and teachers (Sanne, Torsheim, Heiervang, & Stormark, 2009). Another reason for the frequent use of the SDQ is that it is brief and easily available (Goodman, 1999).

The majority of the studies on the SDQ and its ability to detect mental health disorders are based on community samples (e.g. Goodman et al., 2000b; Mellor, 2004; van Roy, Grøholt, Heyerdahl, & Clench-Aas, 2006). Goodman and colleagues (2000b) found that the SDQ could potentially increase the detection of mental health disorders when used in a community screening programme. However, the accuracy measures of a screening instrument may change according to which populations it is applied to. Thus, these results cannot be generalized without further ado from one country to another, or from a community population to a clinical population. At the same time, the use of the SDQ in child and adolescent outpatient clinics is increasingly becoming more common in Norway (Grøholt, Sommerschild, & Garløv, 2009). Consequently, it is important that the SDQ sufficiently recognises disorders in these kinds of outpatient clinics. To our knowledge, limited research has been conducted in clinical populations to validate the diagnostic predictions resulting from the SDQ.

In a study by Goodman and colleagues (2000c), clinical samples in England and Bangladesh were used to examine the accuracy of the diagnostic predictions estimated by the SDQ. They developed a computerised algorithm on the basis of multi-informant SDQ symptom and impact scores. This algorithm identified four broad categories, namely conduct disorder, emotional disorder, hyperactivity disorder as well as 'any psychiatric disorder'. The outcomes from the algorithm were compared with clinical diagnoses assigned by experienced clinicians according to ICD-10 criteria. These clinical ratings served as a 'gold standard'.

Goodman and colleagues' (2000c) reported the sensitivity for the different categories to be ranging from 81 to 90%. The specificity was reported to be ranging from 78 to 84%, except for predictions of conduct disorder in the English sample (47%). Positive and negative predictive values (PPV and NPV) were also measured in this study. For the PPVs there was a wide range of variation, with values from 35% (hyperactivity disorder) to 86% (emotional disorder). The NPVs ranged from 83

to 98%. The authors argue, on the basis of the sensitivity results, that the SDQ algorithm is good at detecting psychiatric disorders. However, the PPVs suggest it might be slightly over-inclusive (Goodman et al., 2000c).

In a study by Mathai and colleagues (2004), the aim was to replicate Goodman and colleagues' (2000c) study in an Australian environment. The sample was clinical and consisted of new referrals to a Child and Adolescent Mental Health Service (CAMHS) in Australia. Similarly to Goodman and colleagues (2000c), they used a computerised algorithm to generate diagnosis from the SDQ into four broad diagnostic categories. Clinicians at the CAMHS assigned diagnoses to the participants according to DSM-IV criteria. A chief investigator placed the clinical diagnoses into the diagnostic categories. An independent clinician also served as a rater and assigned clinical diagnoses after reading case notes. The findings showed a sensitivity of 44% for hyperactivity disorder, 93% for conduct disorder, and 36% for emotional disorder (Mathai et al., 2004).

In the above mentioned studies case notes were used as the clinicians' basis to assign psychiatric diagnoses to the participants. Both Goodman and colleagues (2000c) and Mathai and colleagues (2004) argue for the need in future studies to use standardised psychiatric interviews as the basis for the diagnoses made by clinicians. This might increase the possibility of identifying a better gold standard. In other studies based on community samples, the Developmental and Well-Being Assessment (DAWBA) interview has been used by clinicians as a basis to assign psychiatric diagnoses to the participants (Goodman et al., 2000b; Heiervang et al., 2007). The DAWBA is a comprehensive interview designed to generate psychiatric diagnosis on the basis of information from multiple sources. It focuses on children's and adolescents' symptoms and the related impact of their problems (Goodman, Ford, Richards, Gatward, & Meltzer, 2000a).

The DAWBA is frequently used in Norwegian child and adolescent mental health clinics. The inconvenience of using DAWBA to assess all patients is the length and comprehensiveness of the interview. Goodman and colleagues (2000a) noted that the parent interview in a community sample approximately takes 50 minutes to administer. The interview is rather extensive, and clinicians use a lot of time interpreting, making clinical judgments, and assigning diagnoses on the basis of DAWBA. To make the intake evaluations more efficient, a shorter screening

instrument like the SDQ may be used in the initial intake process to detect patients who need a more comprehensive assessment. As earlier noted a screening instrument may recognize the possibility that a patient might have a disorder and thereafter identify the need for further assessment (Warner, 2004). The DAWBA or other assessment tools can be used in further evaluation. In addition, it can be useful to be aware of the diagnostic predictions from the SDQ in order to address the child to the presumably appropriate team member (Goodman et al., 2000c).

To use the SDQ as the first step in the intake process provides that it is sufficiently accurate in its predictions of actual disorders. To measure the accuracy of a screening instrument a comparison against a gold standard should be made (Greenhalgh, 1997). Sensitivity and specificity are well-known screening efficiency statistics to measure the accuracy of a screening instrument (Akobeng, 2006). PPVs and NPVs are also used to measure screening efficiency. Sensitivity and specificity can be combined into likelihood ratios, which also are statistic measures to summarize the accuracy of a test (Deeks & Altman, 2004). The aim of the present study was to examine the ability of the SDQ to detect mental health disorders among patients referred to a child and adolescent outpatient clinic. This is the first Norwegian study with a clinical sample aiming to examine the accuracy of the SDQ. In this study the above-mentioned statistical measure were used to compare the diagnostic predictions from the SDQ with diagnoses assigned by clinicians on the basis of DAWBA (the gold standard).

Method

Participants

The data is gathered from a child and adolescent mental health outpatient clinic in Tromsø as part of a larger regional research project in the northern part of Norway. The project lasted from the 1st of September 2006 to the 31st of December 2008. The participants in our study were 100 randomly selected referrals to this outpatient clinic. All participants were referred from either their general practitioner or the child welfare authorities. They all agreed to participate in the study. Only participants with a sufficient amount of data allowing diagnostic analyses were included. These 100 participants were selected in order to estimate the inter-rater reliability of the DAWBA interview and were subsequently assessed at the clinic.

The characteristics of the sample are depicted in Table 1. The age range of the participants was from 5 to 17 years and the mean age was 11.35 (*SD* 3.37). The sample consisted of 58% boys and 42% girls. Fifty-seven per cent of the participants were under 13 years old and 43% between 13 and 17 years old. The parent SDQ and DAWBA were available for 93% of the participants. Some of the parent SDQ and DAWBA reports were completed by foster parents (3%) and grandparents (1%). Teacher SDQ and DAWBA were available for 72% of the participants, and the self-report SDQ and DAWBA were completed by 48% of the adolescents. For 16% of the participants the SDQ and DAWBA reports were completed by only one informant.

Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) and Children's Global Assessment Scale (CGAS) (based on the information in DAWBA) were used to evaluate the severity of the participants' problems. The results are depicted in Table 1.

Table 1

Participant Characteristics

	Total (*N*=100)	Boys (*n*=58)	Girls (*n*=42)	*t*
	M (SD)	*M (SD)*	*M (SD)*	
Age	11.35 (SD 3.37)	10.57 (SD 3.25)	12.43 (SD 3.28)	*t* (98) = -2.81, *p* = .01*
HoNOSCA	11.09 (SD 5.27)	11.03 (SD 4.97)	11.17 (SD 5.72)	*t* (98) = -.13, *p* = .90
CGAS	56.11 (SD 10.56)	55.32 (SD 10.99)	57.19 (SD 9.97)	*t* (98) = -.87, *p* = .39

Note. *$p < .05$.

Instruments and measures

The Strengths and Difficulties Questionnaire (SDQ). The SDQ is a screening instrument for assessing the behaviour of children and adolescents aged 4 to 16 (Goodman, 1997). The SDQ can be administered to parents and teachers, as well as children and adolescents aged 11 years or older. The questionnaire consists of 25 different items constituting five different clinical scales: hyperactivity/inattention, emotional symptoms, conduct problems, peer relationship problems and prosocial behaviour. It covers both resources and problems in children and adolescents. The extended version of the SDQ also includes an impact supplement in which the respondents are asked questions concerning the severity of their difficulties, chronicity, overall distress, social impairment, and burden to others (Goodman, 1999). With the use of impact scores as well as symptom scores, Goodman and colleagues (2000c) developed a predictive algorithm which combined the two scores. This algorithm generates ratings of 'unlikely', 'possible', or 'probable' for four broad categories of diagnoses: Conduct disorder, emotional disorder, hyperactivity disorder, as well as an 'any disorder' category. This predictive algorithm was first employed in a study by Goodman and colleagues (2000c), and can be found on www.sdqinfo.com/e8.html.

Results from different studies on the SDQ's reliability indicate good inter-informant reliability with reports of moderate correlations (Goodman, 2001; Mathai et al., 2002; Mellor, 2004). Internal consistency has been reported to be satisfactory in several international studies (Bourdon and colleagues, 2005; Goodman, 2001). Studies have also shown that the SDQ correlates highly with other behavioural screening questionnaires, such as the Child Behaviour Checklist (CBCL) (Goodman & Scott, 1999) and the Rutter questionnaires (Goodman, 1997).

The factor structure of the SDQ has also been explored, but the different studies show mixed results. For example, Percy, McCrystal, and Higgins (2008) and Mellor and Stokes (2007) found limited support for the factor structure. However, other studies have found support for Goodman's five factor model (e.g. Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004; Smedje et al., 1999). A Norwegian study also confirms the factor structure of the SDQ (Sanne et al., 2009). Sanne and colleagues found that a slightly modified version of Goodman's factor

structure fitted the data best. The goodness of fit was reported to be acceptable according to the authors.

The SDQ exists in both a paper-pencil and an electronic version. The electronic version can be found on www.sdqinfo.com. In the present study the authorized electronic Norwegian version included in the DAWBA was used.

Developmental and Well-Being Assessment (DAWBA). The Development and Well-Being Assessment is a detailed semi-structured psychiatric interview. It contains a mixture of closed and open-ended questions about the symptoms and the resulting impact of child and adolescent psychopathology (Goodman et al., 2000a). There are four components involved in the DAWBA; a parent interview, a teacher questionnaire, an interview for adolescents over the age of 11, and clinical diagnostic ratings. These clinical diagnostic ratings were calculated by a computer based on the answers from the interviews and the questionnaire (Goodman et al., 2000a). The DAWBA is designed to generate psychiatric diagnoses on children and adolescents between the age of 5 and 17, based on ICD-10 and DSM-IV criteria (Fleitlich-Bilyk & Goodman, 2004). Problem areas covered by the DAWBA includes anxiety, depression, ADHD, and behaviour problems, as well as less frequent problems such as eating disorders, tics, and autism (Goodman et al., 2000a).

Goodman and colleagues (2000a) presented evidence supporting the validity of the DAWBA, in showing among others a substantial overlap between the diagnoses generated by the DAWBA and diagnoses assigned by clinicians using case notes. In terms of inter-rater reliability, a longitudinal study of child mental health in Norway revealed very high kappa values between the clinical raters over all categories of disorders (.91 - 1.00) (Heiervang et al., 2007).

The DAWBA exists in both a paper-pencil version and an electronic version. In this study the electronic Norwegian version of the instrument was used. Further information on the DAWBA and the interview itself, can be found and downloaded on www.dawba.com.

Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA). Health of the Nation Outcome Scales for Children and Adolescents is a brief measure of emotional and behavioural problems for children and adolescents between 3 and 18 years (Garralda et al., 2000). HoNOSCA consists of 15 items, where the first 13 items cover clinical features concerning the child or adolescent,

and add up to a total score. The two remaining items rate the parents' understanding of the difficulties, as well as information about services. The first 13 items are categorised into four subsections: Behaviour (1-4), impairment (5-6), symptoms (7-9), and social (10-13) (Burgess, Trauer, Coombs, McKay, & Pirkis, 2009). All 15 items are rated on a five-point severity scale, ranging from 'no problem' to 'severe problems' (Bilenberg, 2003). Burgess and colleagues (2009) suggest that a score of 2 on each item can be evidence of clinically significant difficulties that call for further follow-up. Garralda and colleagues (2000) found a mean score of 11.40 (*SD* 4.89) in a British clinical sample. Mean scores from Australian clinical samples has been reported to be 12.60 (*SD* 6.70) (Australian Mental Health Outcomes and Classification Networks, 2005) and 13.11 (*SD* 6.30) (Brann, Coleman, & Luk, 2001). Several studies have aimed to validate the scale and have found good results on both validity and reliability of the instrument (Bilenberg, 2003; Garralda et al., 2000). The usefulness of assessing clinical change and outcome in child and adolescent mental health clinics was also confirmed (Garralda et al., 2000). In the present study only the clinical features (item 1 to 13) were included.

Children's Global Assessment Scale (CGAS). The Children's Global Assessment Scale is a clinician-administered scale which provides a rating of social and psychological functioning for children and adolescents aged 4 to 16 (Schaffer et al., 1983). The scale ranges from 1 to 100, with 1 being 'severe dysfunction' and 100 being 'superior functioning'. Empirically derived cut-off points suggest that scores above 70 indicate normal functioning (Dyrborg et al., 2000). Mean scores in clinical samples have been found to be 53.92 (*SD* 10.90) in a British sample (Garralda et al., 2000) and 58.80 (*SD* 14.10) for an Australian sample (Australian Mental Health Outcomes and Classification Networks, 2005). Studies on the psychometric properties of CGAS have mainly concentrated on concurrent validity and inter-rater reliability, with good results (Pirkis, Burgess, Kirk, Dodson, & Coombs, 2005). Adequate results are seen for other types of reliability and validity (Pirkis et al., 2005), for example has the discriminant validity been established by Shaffer and colleagues (1983) and Steinhausen and Metzke (2001).

Procedure

The present study was approved by the Regional Committee for Medical Research Ethics and the Norwegian Social Science Data Service.

The participants signed a written informed consent before participating in the study. In families where the child was under the age of 12, parents signed the consent form. For participants between 12 and 15 both parents and the adolescent had to consent according to Norwegian legislation (The Health Research Act, 2008). For participants aged 16 or older only the adolescent had to consent. They were all informed about confidentiality and that participation was entirely voluntary. All the participants were given a private access code to the internet-based electronic version of DAWBA in which the SDQ was included. Parents, teachers and adolescents over the age of 11 completed this interview on www.dawba.net.

Four independent clinicians, one child psychiatrist and three neuropsychologists completed an online training programme for scoring the DAWBA. In addition, one of the raters was trained by the developer of DAWBA, Dr. Robert Goodman. To ensure comparable rating thresholds with previous studies using DAWBA, this rater guided the other raters in the following training sessions. All raters in the present study were trained in scoring HoNOSCA and CGAS by the use of vignettes in a one-day training session (Hanssen-Bauer et al., 2006). The raters also attended two training workshops for HoNOSCA and CGAS, which lasted for two days each.

Subsequent to the training sessions, each of the raters rated severity of the reported problems using HoNOSCA and CGAS. The raters also individually rated the 100 participants according to the DSM-IV manual (American Psychiatric Association, 1994). The participants who fulfilled the criteria were given a DSM-IV diagnosis. The various DSM-IV diagnoses were then allocated into three broad diagnostic categories: emotional disorder, hyperactivity disorder and conduct disorder. The category emotional disorder included anxiety disorders and depressive disorders. The category hyperactivity disorder included the different attention deficit hyperactivity disorders (ADHD). The category conduct disorder included oppositional defiant disorder and conduct disorder. The implementation of these categories was adopted from Goodman and colleagues (2000c). The same categories have also been employed by Mathai and colleagues (2004).

After the individual rating, the raters discussed all cases where any of the raters disagreed (n = 25). The cases were discussed until consensus was achieved, and the conclusion was used as the gold standard. Similar procedures have been used

in preceding studies, for example in the Bergen Child Study (Heiervang et al., 2007) and the British Child and Adolescent Mental Health Survey in 1999 (Ford, Goodman, & Meltzer, 2003).

To assess the level of agreement between the four clinicians' diagnostic ratings, the inter-rater reliability was examined (Brøndbo et al., 2010). The kappa coefficients were .81 - 1.00 for conduct disorder, .71- .91 for emotional disorder, .70-.85 for hyperactivity disorder, and .73 - .85 for 'any disorder', ranging from substantial to almost perfect agreement according to Landis & Koch's (1977) categorisation of kappa ranges.

Statistical analyses

Analyses were conducted in SPSS version 16.0 (SPSS Inc., Chicago). Chi Square analyses were conducted to compare boys and girls in relation to the different diagnostic categories. This was done for both the SDQ predictions and the diagnoses made by clinicians. An independent-samples t-test was conducted to compare the mean age for boys and girls. Independent samples t-tests were also conducted for HoNOSCA and CGAS scores to compare the means for boys and girls.

The SDQ diagnostic algorithm. The diagnostic algorithm, as programmed in SPSS version 16.0 (SPSS Inc., Chicago), combined the symptom and impact scores on the SDQ, from parents, teachers and the adolescents. The algorithm in SPSS generated three levels of probability for the four diagnostic categories. The three levels of probability were 'unlikely', 'possible' and 'probable'. To calculate the values for screening efficiency (sensitivity, specificity, PPV, NPV, positive and negative likelihood ratios), we dichotomised the SDQ probability levels into 'diagnosis' and 'no diagnosis' (Goodman et al., 2000b; Goodman et al., 2000c; Mathai et al., 2004). 'Unlikely' and 'possible' predictions of the algorithm were counted as 'no diagnosis'. 'Probable' was counted as 'diagnosis'. This dichotomisation is necessary to calculate the screening efficiency in terms of sensitivity, specificity, PPV and NPV (Goodman et al., 2000c).

Screening efficiency statistics. To measure the screening efficiency of the SDQ the following measures were used: sensitivity, specificity, PPV, NPV, positive likelihood ratio and negative likelihood ratio. The calculations of these measures are based on the information in Table 2.

Table 2

Performance of a Screening Test

	Gold standard		
Screening test	Diagnosis	No diagnosis	Total
Diagnosis	*a*	*b*	*a* + *b*
No diagnosis	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

Table 2 is a 2 x 2 model in which the rows show the results from a screening test and the columns show the results of a gold standard. The letter 'a' represents true positives, 'b' represents false positives, 'c' represents false negatives and 'd' represents true negatives (Greenhalgh, 1997).

Sensitivity and specificity is one way of quantifying the diagnostic accuracy of a test (Altman & Bland, 1994). Sensitivity is the ability of the instrument to generate a positive result for someone with the diagnostic category of interest (Glaros & Klines, 1988). Specificity is the ability of the instrument to generate a negative result for someone without the diagnosis of interest. To calculate sensitivity and specificity the following equations were used:

$$Sensitivity = \frac{a}{a + c}$$

$$Specificity = \frac{d}{b + d}$$

Sensitivity and specificity are important when determining diagnostic accuracy, but is not useful in estimating the probability of a disorder (Akobeng, 2006). Positive and negative predictive values refer to the probability that a positive or negative test result is correctly diagnosed (Glaros & Klines, 1988). These values vary according to the prevalence of a disorder in the population (Warner, 2004). For example, if the prevalence of a disorder is low, the PPV will be low even if the

specificity and sensitivity are very high. To calculate PPV and NPV the following equations were used:

$$PPV = \frac{a}{a+b}$$

$$NPV = \frac{d}{c+d}$$

In the present study the values for sensitivity, specificity, NPV and PPV were converted from decimal numerals to percentages.

Likelihood ratios are ratios of probabilities, and are used to summarise diagnostic accuracy on the basis of sensitivity and specificity (Deeks & Altman, 2004). The likelihood ratio provides information about how a test result (positive or negative) will change the likelihood for someone to have a certain diagnosis. We have interpreted the sizes of the likelihood ratios according to Jaeschke and colleagues' guide (Jaeschke, Guyatt & Sackett, 1994). To calculate positive and negative likelihood ratio the following equations were used:

$$Positive\ likelihood\ ratio = \frac{sensitivity}{1 - specificity}$$

$$Negative\ likelihood\ ratio = \frac{1 - sensitivity}{specificity}$$

Confidence intervals were included so that the values are more precisely displayed (Warner, 2004). An internet-based calculator was used to calculate the confidence intervals for all values in the screening efficiency statistics in our study (Hozo & Djulbegovic, 1999). This JavaScript program is available at www.iun.edu/~matio/medmath/old/ci-java.htm.

Results

There was a significant difference (t (98) = -2.81, p = .01) between the mean age of boys and girls, as shown in Table 1. For HoNOSCA, the mean score for all

participants was 11.09 (*SD* 5.27), and 56.11 (*SD* 10.56) for CGAS. The independent samples t-test showed no significant differences between boys' and girls' mean scores on HoNOSCA and CGAS. The information is also shown in Table 1.

Table 3 illustrates the SDQ-predicted diagnoses. The SDQ predicted that 72% of the participants had a psychiatric disorder which fits into one of the three diagnostic categories. Conduct disorder was the diagnostic category with the largest percentage in total (47%). In terms of gender differences, the percentages for prediction of conduct disorder and hyperactivity disorder were higher for boys than for girls. The difference between boys and girls for predictions of hyperactivity was significant ($\chi^2 = .05$, $p < .05$). On the contrary, a larger percentage of girls was predicted to have an emotional disorder than boys. This difference was significant ($\chi^2 = .03$, $p < .05$). The predictions for having 'any disorder' were almost identical for both boys and girls.

Table 3

Percentage of Psychiatric Disorders Predicted by the SDQ

	Total (*N*=100)	Boys (*n*=58)	Girls (*n*=42)	χ^2
Conduct	47%	50%	33%	.48
Emotional	27%	19%	38%	.03*
Hyperactivity	27%	34%	17%	.05*
Any disorder	72%	71%	74%	.73

Note. *$p < .05$.

Diagnoses assigned by clinicians (the gold standard) are presented in Table 4. The clinicians assigned 70% of the participants a psychiatric disorder which fit into one of the diagnostic categories. This distribution shows a similar pattern of gender differences as in the SDQ-predictions (see Table 3). More girls were assigned an emotional disorder than boys, and a larger percentage of boys were assigned a

conduct or hyperactivity disorder than girls. A significant difference was found between boys and girls for conduct disorder ($\chi^2 = .03$, $p < .05$) and emotional disorder ($\chi^2 = .04$, $p < .05$). However, the gender differences on hyperactivity disorder are less noticeable in the clinical diagnoses than in the SDQ predictions. Similar to the SDQ-predictions, conduct disorder was the diagnostic category with the largest percentage in total (39%), though somewhat lower than the SDQ-predictions (47%).

Table 4

Percentage of Psychiatric Disorders Assigned by Clinicians

	Total (*N*=100)	Boys (*n*=58)	Girls (*n*=42)	χ^2
Conduct	39%	48%	26%	.03*
Emotional	34%	26%	45%	.04*
Hyperactivity	22%	26%	17%	.27
Any disorder	70%	71%	69%	.86

Note. *$p < .05$.

Table 5 presents the screening efficiency for SDQ in terms of sensitivity, specificity, PPV, NPV, positive likelihood ratio and negative likelihood ratio for the different diagnostic categories conduct, emotional, hyperactivity and 'any disorder'. Confidence intervals (95% level) are also included in Table 5. The sensitivity for conduct disorder, hyperactivity disorder and 'any disorder' was between 77% and 90%, while the sensitivity for emotional disorder was considerably lower (56%). The specificity was between 70% and 88%, with emotional disorder and hyperactivity disorder being noticeably higher. The PPV was highest for 'any disorder' with 88%. For the other three diagnostic categories the PPV ranged from 63% to 70%. In terms of the NPVs they varied across all diagnostic categories with percentages from 75% to 93%. A moderate positive likelihood ratio (6.03) was found for hyperactivity

disorder. The positive likelihood ratios for conduct disorder and 'any disorder' were small, respectively 3.03 and 3.00. For emotional disorder the positive likelihood ratio was small but close to moderate (4.61). The negative likelihood ratio for 'any disorder' was 0.14, which is moderate. For the other three categories the negative likelihood ratios were small, ranging from 0.26 to 0.50.

Table 5
Screening Efficiency for the Diagnostic Categories

	Conduct		Emotional		Hyperactivity		Any disorder	
	Scores	95% CI	Scores	95% CI	Scores	95% CI	Scores	95% CI
Sensitivity	79%	[63, 90]	56%	[38, 72]	77%	[54, 91]	90%	[80, 96]
Specificity	74%	[61, 84]	88%	[77, 94]	87%	[77, 93]	70%	[50, 85]
Positive predictive value	66%	[51, 79]	70%	[50, 86]	63%	[42, 80]	88%	[77, 94]
Negative predictive value	85%	[72, 93]	79%	[68, 88]	93%	[84, 97]	75%	[55, 89]
Positive likelihood ratio	3.03	[1.93, 4.75]	4.61	[2.26, 9.42]	6.03	[3.24, 11.22]	3.00	[1.73, 5.21]
Negative likelihood ratio	0.28	[0.15, 0.52]	0.50	[0.34, 0.74]	0.26	[0.12, 0.57]	0.14	[0.07, 0.30]

Note. CI = confidence interval.

Discussion

The aim of the present study was to examine the ability of the SDQ to detect mental disorders among patients referred to a child and adolescent outpatient clinic. We examined this by comparing the outcome of diagnoses assigned by the experienced clinicians (the gold standard), with the diagnoses generated by the SDQ algorithm. The results show that 72% of the children and adolescents in the sample have a diagnosable mental disorder according to the SDQ. The prevalence of emotional disorder in this sample was 27%. The prevalence of hyperactivity disorder was also 27%, while the rate for conduct disorder was substantially higher (47%).

The results of our study demonstrated that our sample was quite similar to samples in comparable clinical studies from other countries. When we compared the ratings of the severity of the participants' problems with HoNOSCA and CGAS, the ratings were similar across samples from both Australia (Australian Mental Health Outcomes and Classification Networks, 2005; Brann et al., 2001) and England (Garralda et al., 2000). Findings from Australia displayed mean scores of HoNOSCA of 12.60 (*SD* 6.70) and mean scores of CGAS of 58.80 (*SD* 14.10) (Australian Mental Health Outcomes and Classification Networks, 2005). Brann and colleagues (2001) found a mean score of 13.11 (*SD* 6.30) for HoNOSCA. The British findings for HoNOSCA were 11.40 (*SD* 4.89) and 53.92 (*SD* 10.90) for CGAS (Garralda et al., 2000). The results from the present study are very similar to the above-mentioned reports, with 11.09 (*SD* 5.27) for HoNOSCA and 56.11 (*SD* 10.56) for CGAS, which suggests that the clinical sample in this study is comparable to other countries.

Our sample also holds several of the characteristics of an average population in a Norwegian child and adolescent outpatient mental health clinic. The most frequently assigned diagnoses in Norwegian child and adolescent outpatient clinics fall within the categories hyperactivity disorder, emotional disorder and conduct disorder (Andersson, 2009). This implies that the SDQ has the ability to identify a broad scope of disorders common in a Norwegian clinical population.

In terms of gender and age distributions our results demonstrate that the difference in mean age between boys and girls in this sample was statistically significant. There were more girls between the age of 13 and 17 years. On the other hand, there were more boys under the age of 13 than girls. The same pattern of age distribution has been reported from Norwegian child and adolescent mental health

services (Andersson, 2009). The results in the present study show that more boys than girls are diagnosed (according to the SDQ) with hyperactivity and conduct disorders than girls. The difference between boys and girls was statistically significant for hyperactivity disorder. The opposite pattern was demonstrated in terms of emotional disorder, with more girls being diagnosed. This difference was also statistically significant. These results are in accordance with the results from a Norwegian report of the child and adolescent mental health services (Andersson et al., 2005).

The SDQ screening efficiency was measured for each diagnostic category, as well as for the presence or absence of 'any diagnosis'. The results for conduct disorder demonstrated a fairly high sensitivity (79%). This means that the SDQ correctly identified 79% of the patients with a conduct disorder. Mathai and colleagues (2004) reported a sensitivity of 93% for conduct disorder in their study, while Goodman and colleagues (2000c) reported a sensitivity of 90% (English sample) and of 86% (Bangladeshi sample). These comparable studies showed slightly better percentages than our study. The PPV for conduct disorder in the present study was low (66%). Thus, the chance that the diagnosis is correct is small. Similar percentages were reported by Goodman and colleagues with 68% and 60%. The PPVs in our and the above-mentioned study indicate that the SDQ prediction for conduct disorder was slightly over-inclusive, which means that there were a somewhat high percentage of false positives. Goodman and colleagues found the specificity for conduct disorder to be 47% and 82% for the two samples, while in this study the specificity was 74%. The specificity for conduct disorder was the lowest across the diagnostic categories.

The positive likelihood ratio for conduct disorder was small (3.03). This means that it is about three times more likely that patients with a conduct disorder will have a positive test result on the SDQ, than patients without conduct disorder. The negative likelihood ratio was also small (0.28), meaning that it was about three to four times more likely for patients without conduct disorder to have a negative test result on the SDQ, than patients with conduct disorder. The overall results imply that the SDQ is neither good enough to identify nor to rule out conduct disorder in a clinical population.

The sensitivity of emotional disorder was only 56%. This means that the SDQ correctly detected just a little more than half of the patients who actually had an emotional disorder. It also implies that the number of false negatives was high, as is additionally shown by the NPV (79%). In a clinical population it is expected that a screening test should be able to detect more than half of the patients with an emotional disorder. Mathai and colleagues' (2004) results displayed a sensitivity of only 36% for emotional disorder. The sensitivity of emotional disorder in Goodman and colleagues' (2000c) results was also somewhat lower than the other diagnostic categories in their study. It is possible that this pattern in both our study and the aforementioned studies, reflects the idea that internalizing behaviour is hard to detect for the surrounding environment. In the present study there were more reports on the SDQ from parents and teachers than self-reports. Parents and teachers may have more difficulties identifying symptoms of an emotional disorder than the child or adolescent itself. In addition, the criteria for emotional disorder are based on a person's own experiences.

The NPV for emotional disorder in the present study (79%) was substantially lower than Goodman and colleagues' (2000c) NPVs, 94% and 90% respectively for England and Bangladesh. We were surprised to learn that even though the sensitivity for emotional disorder was low, the positive likelihood ratio for emotional disorder was small, but close to moderate (4.61). On the other hand, the negative likelihood ratio was small, but close to tiny (0.50). This means that it is only twice as likely that patients without an emotional disorder have a negative test result on the SDQ, than patients with an emotional disorder. A result giving so many false negatives implies that many patients with an emotional disorder might not get the help they need.

The results displayed a sensitivity of 77% for the category hyperactivity disorder. Comparing this to Goodman and colleagues (2000c), their result showed a sensitivity of 89% for both samples. Mathai and colleagues (2004) had a considerably lower percentage of 44%. Both the specificity and the NPV for hyperactivity disorder in the present study displayed high percentages (87% and 93%), thus the SDQ correctly identified a large amount of the patients without the disorder. The results for the same measures in Goodman and colleagues' study, showed equivalently high values of NPVs (93 - 98%), but slightly lower for the

specificity (78 - 81%) than in the present study. The PPV was low (63%), indicating that the SDQ predictions for hyperactivity was slightly over-inclusive.

For hyperactivity disorder the positive likelihood ratio was moderate (6.03), and it was the highest positive likelihood ratio in this study. This means that it is about six times more likely that patients with a hyperactivity disorder will have a positive test result on the SDQ, than patients without the disorder. This result indicates that the SDQ is somewhat useful in screening for hyperactivity disorder. The negative likelihood ratio for the same disorder was small (0.26), indicating that patients without hyperactivity disorder is four times more likely to have a negative test result, than patients with the disorder. The results for the diagnostic categories demonstrates that the SDQ predictions for hyperactivity disorder were better than the other categories to correctly identify true negatives, which is an important aspect of a screening instrument. Correctly identifying many true negatives means that there will be few false negatives. However, it might not be good enough to use as a screening instrument for hyperactivity disorder in a general mental health clinic, but perhaps it could be useful to aid clinicians in a more specialised hyperactivity clinic.

The present study also included the category 'any disorder'. The SDQ identified any psychiatric disorder in 72% of the sample in the present study. In Goodman and colleagues' (2000b) community study the corresponding number for any psychiatric disorder was 10.5%. This difference in prevalence highlights the difficulty with generalising PPV results from community populations to clinical population, hence the importance of the present study.

Screening efficiency statistics were also conducted for the category 'any disorder'. Neither Goodman and colleagues (2000c) nor Mathai and colleagues (2004) conducted screening efficiency statistics for this category. In the present study, the sensitivity for 'any disorder' had the highest percentage (90%) of sensitivity out of all diagnostic categories. The PPV also displayed the highest percentage (88%) when comparing the different categories. A high PPV means that a large proportion of the patients with a positive test result are likely to have any psychiatric disorder. The PPV changes according to the prevalence in the population (Altman & Bland, 1994). Therefore, it is important for a clinic to be aware of the prevalence of different disorders in their population. It is a premise that one has knowledge about the population in question so the clinic can be able to screen for

psychiatric disorder. According to Miles and Gilbert (2005) it is important that the PPV is high to assess the costs and benefits of following up the screening test with further evaluation. In other words, one wants to be sure that those who are identified with a diagnosis actually have a disorder before resources are being used for these patients. The NPV for any disorder was 75%. This means that the SDQ generated a large number of false negatives.

The positive likelihood ratio for any disorder was small (3.00), and was the lowest values across the diagnostic categories. According to the moderate negative likelihood ratio (0.14) for any disorder it is about seven times more likely for patients without any disorder to have a negative test result, than patients with any disorder. This negative likelihood ratio was the best across the categories. In conclusion, the SDQ is good at identifying any psychiatric disorder, but the low result for NPV raises the question of how accurate the SDQ is at ruling out patients without any psychiatric disorder.

One of the practical advantages of the SDQ is its use of multiple informants (Goodman, 1997). Goodman and colleagues (2000b) found that the best result for the SDQ prediction in a community sample was when all possible informants had completed the SDQ. In the present study, the number of informants for each participant varied. For 16% of the participants the SDQ reports and DAWBA were completed by only one informant. Ideally, one should have had multiple informants for the whole sample. However, this is a clinical sample and the number of informants is not possible to control. For example, adolescents over the age of 16 referred to mental health clinics may not want to have their parents involved in the assessment and treatment. Despite this, the information from one informant is still interesting and SDQ predictions can be generated.

To implement the SDQ as a first step in the intake process at a clinic, one would have to consider which qualities of the test are considered most important. It is not possible to have a test that is 100% accurate in its decision of whether a person has a disorder or not, or what type of disorder this person has (Bhopal, 2008). In the context of an early evaluation process in clinics, for example a high specificity for the category hyperactivity disorder means that the SDQ detects a large percentage of those without hyperactivity disorder. It is useful to correctly identify those who do not have a disorder early in an intake process, as for one reason many outpatient

clinics already have problems with the capacity. This way, the clinic might have more resources for other patients who actually do have a disorder. However, if the sensitivity at the same time is low, there will be a larger amount of patients with a false negative result (patients in which the SDQ has predicted no diagnosis, but who actually have a disorder). This may lead to some patients with such a result to be rejected from the clinic, leaving these patients who actually have a hyperactivity disorder without a proper follow-up.

On the other hand, if the sensitivity is high like in 'any disorder' (90%), and the specificity is a bit lower (70%), there is a risk of false positives (patients in which the SDQ has predicted a diagnosis, but who actually do not have a disorder). The consequence of these patients getting a false positive result might be to cause more distress and concern in patients as well as in their parents. These patients have to be the object for further comprehensive assessments, which can lead to resources at the clinic being used at the expense of other patients with an actual disorder. Though these false positives may create more concern in the family, it is a good chance that the families are already quite worried since the patient is referred to an outpatient clinic. This could possibly lead to a worse consequence in community samples. When screening the referrals in an early evaluation process in a mental health clinic there are also ethical considerations to bear in mind. If patients with a disorder are rejected from getting help, that might be a larger ethical concern than using somewhat extra resources to further the assessment on patients without a disorder. With these ethical considerations in mind we conclude that it is more problematic if a screening instrument in a clinical population rejects a lot of false negatives, rather than risking some false positives.

There are several limitations to our study. One limitation is that the raters were not blind to the SDQ predictions while assigning the clinical diagnoses based on DAWBA. The reason for this was that the SDQ was included in DAWBA. This might have affected the raters' clinical assessment. For example, if the SDQ prediction for a participant was an emotional disorder, the raters could have been primed to assign an emotional disorder to this participant. In similar studies the clinical raters have been blind to the SDQ scores to avoid this bias (e.g. Mathai et al., 2004; Goodman et al., 2000b; Goodman et al., 2000c). However, it can be considered

more ecologically valid since the SDQ is usually included in DAWBA when used in outpatient clinics.

Assuming that the clinical diagnosis is the gold standard also has its limitations. Mathai and colleagues (2004) noted the importance of taking this into consideration when evaluating the results. In the present study a standardised interview were used as the basis for clinicians to assign diagnoses, as recommended by Goodman and colleagues (2000c) and Mathai and colleagues (2004). The gold standard can never be 100% accurate, but on the other hand, to date there is no better gold standard to apply.

Lastly, another limitation is the size of the sample which did not allow us to examine gender differences within the specific diagnostic categories. In a further study with a larger clinical sample size it could be interesting to examine the differences in the SDQ predictions and examine which informants are the best predictors of psychiatric disorders, based on type of informants and number of informants. With a larger sample size it would also be interesting to examine other subgroups, for example gender, age, various risk factors and resilience factors.

Warnick et al. (2008) stated that a great deal of the research on the SDQ has been based on European and Australian populations. It will thus be important to conduct more research in other parts of the world.

It could also be interesting in future studies to use raters who have not undergone training in using different instruments like DAWBA, HoNOSCA and CGAS. This scenario is much more common in a daily clinical practice, and Lundh, Kowalski, Sundberg, Gumpert, and Landén (2010) found a large difference between trained and untrained raters in the rating of different scales.

Research on screening tests in the field of psychology is not as straight forward as research in the medical field. Since there are no blood samples or biological markers revealing psychiatric disorders at hands, it is of utmost importance that the screening instruments and diagnostic tests available are as useful and accurate as possible. The overall results of the present study demonstrated that the SDQ was sufficiently able to detect the presence of any mental health disorder. However, the correct detection of true negatives for any disorder may not be accurate enough for the SDQ to be used as the first step in the intake process, because of the risk of rejecting those who actually need help. Consequently, the clinic would have

to use further assessments regardless of the SDQ results. In terms of type of disorder, the SDQ showed somewhat better accuracy in predicting hyperactivity disorder than the other diagnostic categories. Though the accuracy was better for hyperactivity than the other categories, the SDQ is not sufficiently accurate in the detection of hyperactivity disorder to be used in screening in a clinical population.

In conclusion, the results from the present study indicate that the SDQ might not be accurately enough to be of practical use as a screening instrument in a child and adolescent mental health clinic.

References

Act on Medical and Health Research (the Health Research Act): ACT 2008 – 06 – 20 no. 44. (2008).

Akobeng, A. K. (2006). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Pædiatrica, 96,* 338-341.

Altman, D. G., & Bland, J. M. (1994). Statistic notes. Diagnostic tests 1: Sensitivity and specificity. *British Medical Journal, 308,* 1552.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Health Disorders* (4th ed.). Washington, DC: Author.

Andersson, H. W. (2009). Pasienter og behandlingstilbud i psykisk helsevern for barn og unge (Report No. SINTEF A9714). Trondheim: SINTEF Teknologi og samfunn - Helsetjenesteforskning.

Andersson, H. W., Ose, S. O., & Sitter, M. (2005). Psykisk helsevern for barn og unge - behandlernes og brukernes vurdering av behandlingstilbudet (Report No. STF78 A055009). Trondheim: SINTEF Helse.

Australian Mental Health Outcomes and Classification Networks. (2005). Child & adolescent national outcomes & casemix collection standard reports (1st. ed) (Report No. 1.1). Brisbane: Author.

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* (Supplement 2), ii11-ii16.

Bilenberg, N. (2003). Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA). Results of a Danish field trial. *European Child and Adolescent Psychiatry, 12,* 289-302.

Bhopal, R. S. (2008). *The concepts of epidemiology. Integrating the ideas, theories, principles and methods of epidemiology* (2nd ed.). New York: Oxford University press.

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 American Academy Child and Adolescent Psychiatry, 44,* 557-564.

Brann, P., Coleman, G., & Luk, E. (2001). Routine outcome measurement in a child and adolescent mental health service: An evaluation of HoNOSCA. *Australian and New Zealand Journal of Psychiatry, 35,* 370-376.

Brøndbo, P. H., Mathiassen B. I., Eriksen, M., Heiervang E., Kvernmo, S., & Martinussen, M. (2010). *Diagnostic agreement between clinicians in child and adolescent mental health service.* Manuscript in preparation.

Burgess, P., Trauer, T., Coombs, T., McKay, R., & Pirkis, J. (2009). What does 'clinical significance' mean in the context of the Health of the Nation Outcome Scale. *Australasian Psychiatry, 17,* 141-148.

Deeks, J. J., & Altman, D. G. (2004). Statistics notes. Diagnostic tests 4: Likelihood ratios. *British Medical Journal, 329,* 168-169.

Dyrborg, J., Larsen, F. W., Nielsen, S., Byman, J., Nielsen, B. B., & Gautrè-Delay, F. (2000). The Children's Global Assessment Scale (CGAS) and Global Assessment of Psychosocial Disability (GAPD) in clinical practice - substance and reliability as judged by intraclass correlations. *European Child and Adolescent Psychiatry, 9,* 195-201.

Fleitlich-Bilyk, B., & Goodman, R. (2004). Prevalence of child and adolescent psychiatric disorders in Southeast Brazil. *Journal of American Academy Child and Adolescent Psychiatry, 43,* 727-734.

Ford, T., Goodman, R., & Meltzer, H. (2003). The British Child and Adolescent Mental Health Survey 1999: The prevalence of DSM-IV disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 1203-1211.

Garralda, M. E., Yates, P., & Higginson, I. (2000). Child and adolescent mental health service use. HoNOSCA as an outcome measure. *British Journal of Psychiatry, 177,* 52-58.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44,* 1013-1023.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38,* 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, 40,* 791-799.

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, 40,* 1337-1345.

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000a). The Development and Well-Being Assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry, 41,* 645-655.

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000b). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry, 177,* 534-539.

Goodman, R., Renfrew, D., & Mullick, M. (2000c). 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,* 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,* 17-24.

Greenhalgh, T. (1997). How to read a paper: Papers that report diagnostic or screening tests. *British Medical Journal, 315,* 540-543.

Grøholt, B., Sommerschild, H., & Garløv, I. (2008). *Lærebok i barnepsykiatri* (4th ed.). Oslo: Universitetsforlaget.

Hanssen-Bauer, K., Gowers, S., Aalen, O. O., Bilenberg, N., Brann, P., Garralda, E., . . .Heyerdahl, S. (2007). Cross-national reliability of clinician-rated outcome measures in child and adolescent mental health services. *Administration and Policy in Mental Health and Mental Health Services Research, 34,* 513-518.

Heiervang, E., Stormark, K. M., Lundervold, A. J., Heimann, M., Goodman, R., Posserud, M-J., . . .Gillberg, C. (2007). Psychiatric disorders in Norwegian 8- to 10-year-olds: An epidemiological survey of prevalence, risk factors and service use. *Journal of American Academy of Child and Adolescent Psychiatry, 46,* 438-447.

Helsedirektoratet. (2008). Veileder for poliklinikker i psykisk helsevern for barn og unge (Report No. IS-1570). Retrieved from http://www.helsedirektoratet.no /vp/multimedia/archive/00070/Psykisk_helsevern_fo_70789a.pdf

Heyerdahl, S. (2003). SDQ - Strength and Difficulties Questionnaire: En orientering om et nytt spørreskjema for kartlegging av mental helse hos barn og unge, brukt i UNGHUBRO, OPPHED og TROFINN. *Norsk Epidemiologi, 13,* 127 135.

Hozo, I., & Djulbegovic, B. (1999). Using the Internet to calculate clinical action thresholds. *Computers and Biomedical Research, 32,* 168-185.

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). User's guide to the medical literature. III. How to use and article about a diagnostic test. B. What are the results and will they help me in caring for my patients. *The Journal of the American Medical Association, 271,* 703-707.

Landis, J. R., & Koch, G. G. (1977). The measurements of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lundh, A., Kowalski, J., Sundberg, C. J., Gumpert, C., & Landén, M. (2010). Children's Global Assessment Scale (CGAS) in a naturalistic clinical setting: Inter-rater reliability and comparison with expert ratings. *Psychiatric Research, 177,* 206-210.

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). *Australian e-Journal for the Advancement of Mental Health, 1.* Retrieved from: www.ausienet.com/ journal/vol1iss3/Mathai.pdf.

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,* 639-643.

Mathiesen, K. S. (2009). Psykiske lidelser i Norge: Et folkehelseperspektiv. Del 2: Barn og unge (Report No. 2009:8). Oslo: Nasjonalt folkehelseinstitutt.

Mellor, D. (2004). Furthering the use of the Strengths and Difficulties Questionnaire: Reliability with younger child respondents. *Psychological Assessment, 16,* 396-401.

Mellor, M., & Stokes, M. (2007). The factor structure of the Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment, 23,* 105-112.

Miles, J., & Gilbert, P. (2005). *A handbook of research methods in clinical and health psychology*. Oxford: Oxford University Press.

Obel, C., Heiervang, E., Rodriguez, A., Heyerdahl, S., Smedje, H., Sourander, A., . . .Olsen, J. (2004). The Strengths and Difficulties Questionnaire in the Nordic countries. *European Child & Adolescent Psychiatry, 13* (supplement 2), ii32-ii39.

Percy, A., McCrystal P., & Higgins, K. (2008). Confirmatory factor analysis of the adolescent self-report Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment, 24,* 43-48.

Pirkis, J., Burgess, P., Kirk, P., Dodson, S., & Coombs, T. (2005). Review of standardised measures used in the National Outcomes and Casemix Collection (NOCC). Retrieved from http://amhocn.org/static/files/assets/a83ff69a/Review_of_NOCC_Measures_Version_1.2.pdf

van Roy, B., Grøholt, B., Heyerdahl, S., & Clench-Aas, J. (2006). Self-reported strengths and difficulties in a large Norwegian population 10-19 years: Age and gender specific results of the extended SDQ-questionnaire. *European Child and Adolescent Psychiatry, 15,* 189-198.

Rønning, 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. *European Child and Adolescent Psychiatry, 13,* 73-82.

Sanne, B., Torsheim, T., Heiervang, E., & Stormark, K. M. (2009). The Strengths and Difficulties Questionnaire in the Bergen Child Study: A conceptually and methodically motivated structural analysis. *Psychological Assessment, 21,* 352-364.

Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., & Aluwahlia, S. (1983). A children's Global Assessment Scale (CGAS). *Archives of General Psychiatry, 40,* 1228-1231.

Smedje, H., Broman J.-E., Hetta, E. & van Knorring, A.-L. (1999). Psychometric properties of a Swedish version of the "Strengths and Diffculties Questionnaire". *European Child and Adolescent Psychiatry, 8,* 63-70.

Steinhausen, H. C., & Metzke, C. W. (2001). Global measures of impairment in children and adolescents: Results from a Swiss communitysurvey. *Australian and New Zealand Journal of Psychiatry, 35,* 282-286.

Warner, J. (2004). Clinicians' guide to evaluating diagnostic and screening tests in psychiatry. *Advances in Psychiatric Treatment, 10,* 446-454.

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening efficiency of the Child Behavior Checklist and Strengths and Difficulties Questionnaire: A systematic review. *Child and Adolescent Mental Health, 13,* 140-147.

Woerner, W., Fleitlich-Bilyk, B., Martinussen, R., Fletcher, J., Cucchiaro, G., Dalgalarrondo, P., . . . Tannock, R. (2004). The Strengths and Difficulties Questionnaire overseas: evaluation and applications of the SDQ beyond Europe. *European Child & Adolescent Psychiatry, 13* (Supplement 2), ii47-ii54.

World Health Organization (WHO). (1996). *ICD-10 Classification of mental and behavioural disorders. Multiaxial classification of child and adolescent psychiatric disorders*. Cambridge: Cambridge University Press.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgments project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341-382.