Measurement quality of the Strengths and Difficulties Questionnaire for assessing psychosocial behaviour among Dutch adolescents

Jorien Vugteveen

Cover design: Jorien Vugteveen
Lay out: Douwe Oppewal
Printed by: Ipskamp Printing

The research presented in this thesis was supported by a grant from ZonMw – the Netherlands Organization for Health Research and Development (grant nr. 729300105) and was approved by the ethics committee of the Heymans Institute for Psychological Research of the University of Groningen in the Netherlands (code: 16025-O)
The content of this thesis is partly based on data gathered as part of a study on comparing the validity of the Dutch SDQ and the KIVPA for screening in Dutch child and adolescent social care (ZonMw grant nr. 15700.1021) and the TAKECARE cohort study on tracing achievements, key processes and efforts in professional care for children and adolescents (ZonMw grant nr. 15900.0001)

rijksuniversiteit
groningen

Measurement quality of the Strengths and Difficulties Questionnaire for assessing psychosocial behaviour among Dutch adolescents

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. C. Wijmenga
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 19 november 2020 om 16.15 uur

door

Jorien Vugteveen

geboren op 6 januari 1988
te Hoogeveen

Promotor
Prof. dr. M.E. Timmerman

Copromotor
Dr. A. de Bildt

Beoordelingscommissie
Prof. dr. M.H. Nauta
Prof. dr. P.J. Hoekstra
Prof. dr. F.J. Oort

CONTENTS

1

General introduction

GENERAL INTRODUCTION

Approximately 15 to 25 percent of all adolescents experience psychiatric problems, such as emotional and hyperactivity problems (Fergusson, Horwood, & Lynskey, 1993; Ormel et al., 2015). Early and accurate detection of these problems allows for timely intervention and appropriate monitoring. Detection typically occurs in one of two professional settings. The first is a community setting in which the aim is to identify adolescents at risk of psychiatric disorders among large groups of mainly healthy adolescents. This, for instance, happens during general health check-ups at schools. The second setting is a clinical setting in which the aim is to identify adolescents at risk of psychiatric disorders among adolescents with a wide range of types and severity of psychosocial problems. Moreover, in this setting the aim is to help adolescents by, amongst other things, accurately diagnosing their disorder(s) and describing the difficulties the adolescent encounters. For these purposes, healthcare professionals need information about an adolescent's psychosocial behaviour, preferably obtained from multiple informants (American Psychiatric Association, 2013). One tool that can be used in this process is the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997; Goodman, 1999).

The SDQ is one of the most widely used instruments in screening and diagnostic procedures. It was developed to measure strengths (prosocial behaviour) as well as four types of frequently occurring difficulties (emotional problems, conduct problems, hyperactivity/inattention, and social problems). Additionally, the SDQ aims to measure the impact of psychosocial problems (i.e., the chronicity, distress, social impairment for the adolescent, and burden for others) among adolescents who experience such problems. Each completed SDQ results in a total of seven scale scores: one strengths score, four difficulty scores, one total difficulties score which is the aggregate of the four difficulty scales, and one impact score. Additionally, an externalizing difficulties (the aggregate of the conduct and hyperactivity / inattention difficulties scales) and an internalizing difficulties (the aggregate of the emotional and social difficulties scales) scale score can be calculated.

Validity

An individual's SDQ scale scores, in combination with other sources of information, could give reason for action. The action can for instance pertain to referral to mental healthcare or planning diagnostic procedures. As an adolescent's mental well-being possibly depends on these actions, it is important that the interpretation of the scale scores is substantiated by evidence for their validity (Hubley & Zumbo, 2011; Messick, 1989). Validity is commonly referred to as the degree to which a test measures what it aims to measure (Kelley, 1927). Because this is a rather general description, it is useful to distinguish four types of evidence for validity (Evers et al., 1988). These types refer to the degree to which a test:

1. is subjectively regarded to cover the construct(s) it intends to measure (face validity).
2. is objectively regarded to fully cover all aspects of the construct(s) it intends to measure (content validity).
3. measures the construct(s) it aims to measure (construct validity).
4. results in scale scores that are related to relevant outcomes (criterion validity).

The combined evidence regarding these four types of validity provides an indication of the extent to which the intended interpretation of the test scores is appropriate. As the SDQ was explicitly designed with the DSM-IV (American Psychiatric Association, 1994) and ICD-10 (World Health Organization, 1992) criteria for a select number of disorders in mind (Goodman, 1997; Goodman, Meltzer, & Bailey, 1998; Goodman & Scott, 1999) and the questionnaire has been accepted by mental healthcare professionals and researchers from all around the world, I deem the SDQ's face and content validity to be sufficient. This leaves the two remaining types of validity evidence, construct and criterion validity, to be investigated.

Construct validity. Evidence for construct validity is typically gathered by assessing a standard set of three aspects (Evers, Lucassen, Meijer, & Sijtsma, 2009): 1) a test's presumed internal structure, 2) whether known group differences on the construct(s) measured are indeed reflected in the group's observed test scores, and 3) a test's comparability to other tests that are supposed to measure similar constructs.

The SDQ's presumed internal structure pertains to the five scales (one for strengths and four for difficulties), each measuring one dimension of psychosocial behaviour with five items. Finding support for the SDQ's presumed internal structure would indicate that the domains measured by the five scales can be distinguished from each other, that each scale measures one domain of psychosocial behaviour, and that all items per scale contribute to measuring that domain. Evidence regarding an instruments scale structure is typically gathered through conducting factor analysis (Asparouhov & Muthén, 2009; Evers, Sijtsma, Lucassen, & Meijer, 2010; Muthén, 1984). In case of the SDQ, the analysis needs to be performed for the community and the clinical setting separately, as the validity of scale scores may be context dependent.

Additional information can be gathered by assessing the SDQ's measurement invariance across the community and clinical settings. Measurement invariance implies that SDQ scores gathered in both settings bear the same meaning and can thus be compared to each other. Comparability across settings is essential, as SDQ scores, regardless of the setting they were gathered in, are typically interpreted using community-based norm scores. Evidence regarding measurement invariance across settings can be gathered by conducting a set of multiple-group confirmatory factor analyses (Millsap & Yun-Tein, 2004).

The second construct validity aspect pertains to known groups differences. The SDQ is used in community and clinical settings among adolescents with psychiatric disorders and those without. Compared to adolescents without psychiatric disorders, it is expected that for adolescents with such disorders higher levels of difficulties and weaker levels of prosocial skills are reported. Evidence regarding this construct validity aspect can be gathered by comparing mean scale scores across these groups (Evers et al., 2010).

The third aspect involves the degree to which the SDQ compares to another test that is supposed to measure similar constructs (i.e., convergent validity evidence). This can be assessed through computing correlations between the scales of these tests (Evers et al., 2010). If the SDQ measures what it aims to measure and the other instrument does too, the scale scores of these tests should be strongly related to each other. Additional information can be obtained by comparing the SDQ to another instrument that is supposed to measure different constructs, such as intelligence or development (i.e., discriminant validity evidence). Scale scores of such an instrument and those of the SDQ should be largely unrelated to each other.

Information about the construct validity of the SDQ scales refers to the extent to which each SDQ scale score reflects the specific dimension of psychosocial behaviour that it is presumed to measure. This information helps to understand why the SDQ could be useful in screening and diagnostic procedures, because an instrument that measures what it aims to measure is more likely to be useful for its intended purposes than an instrument that does not. An instrument's value for its intended purposes is a matter of criterion validity.

Criterion validity. Criterion validity refers to the degree to which a test's scale scores are related to relevant external outcomes. The outcomes considered to gather evidence concerning this type of validity highly depend on an instrument's purpose (Evers et al., 2010). As the SDQ is used in screening and diagnostic procedures, evidence regarding its criterion validity must be gathered by investigating its value for use in these procedures. Herewith one needs to focus on identifying adolescents at risk of psychiatric disorders that are related to the domains of psychosocial behaviour covered by the SDQ: Anxiety/Mood disorder, Conduct/Oppositional Defiant Disorder (CD/ODD), Attention-Deficit/Hyperactivity Disorder (ADHD), and Autism Spectrum Disorder (ASD).

Information on how useful the SDQ is for identifying adolescents at risk of psychiatric disorders in a community setting can be obtained by investigating the SDQ's ability to distinguish between adolescents suffering from *any* of the above mentioned psychiatric disorders and adolescents that do not. Additional information can be gathered by assessing *per* disorder (Anxiety/Mood disorder, CD/ODD, ADHD, ASD) how well adolescents suffering from that disorder can be distinguished from adolescents that do not suffer from any of these disorders. The value of the separate SDQ scales for these purposes can be assessed using a receiver operating characteristic curve (Hanley &

McNeil, 1982; Metz, 1978) per scale. Jointly considering all SDQ scales for this purpose is possible using cluster analysis (e.g., Hennig, Meila, Murtagh, & Rocci, 2015), therewith investigating whether SDQ score profiles differ among the above mentioned groups of adolescents.

Compared to adolescents in a community setting, a much larger part of the adolescents in a clinical setting suffers from psychiatric disorders. Moreover, a substantial number of these adolescents likely suffer from more than one disorder, given the high comorbidity rates of psychiatric disorders among youth (Merikangas et al., 2010). Therefore, it is only marginally relevant to investigate in a clinical setting how well adolescents suffering from Anxiety/Mood disorder, CD/ODD, ADHD, or ASD can be distinguished from each other based on SDQ scores. Instead, an indication of the SDQ's value for identifying adolescents at risk of these psychiatric disorders can be obtained by assessing the extent to which SDQ scores are predictive for each of the separate disorders. Per disorder, the predictive ability of single SDQ scales can be assessed using logistic regression analysis. This approach can be extended to simultaneously including all SDQ scales and the four types of psychiatric disorders (and combinations thereof). Another way to jointly consider all SDQ scales and explicitly take into account the potential comorbidity of disorders is by using cluster analysis (e.g., Hennig et al., 2015), therewith investigating whether SDQ score profiles differ among adolescents with different (combinations of) disorders, and content-wise match the specific disorder(s) present among adolescents. For example, among adolescents suffering from both CD/ODD and ADHD a matching SDQ score profile would include high levels of conduct and hyperactivity / inattention difficulties and low scores on the three remaining scales.

Note that the above described types of evidence for the construct and criterion validity should not be viewed as separate, and possibly substitutable, pieces to the puzzle. Instead, they should be considered collectively and in relation to each other for obtaining an indication of the validity of the SDQ score interpretations.

Research aims and thesis outline

Research on validity aspects of the SDQ focusing on Dutch adolescents is scarce. As a result, little is known about how healthcare professionals in the Netherlands should interpret SDQ scores and how useful, if at all, the scores are for the SDQ's intended purposes among adolescents. In order to inform child and adolescent healthcare practice, the studies in this thesis are aimed at gathering evidence regarding construct and criterion validity aspects of the self-report and the parent-report SDQ versions for use in screening and diagnostic procedures among Dutch adolescents aged 12 to 17 years. Additionally, relative norms for interpreting scale scores of both SDQ versions are provided.

Data. This section contains a short description of the data used in the search for evidence regarding the validity aspects. This description helps understand the information provided in the final part of this paragraph, which presents the research aims of the studies described in Chapters two to six of this thesis.

Self-reported and parent-reported SDQ data were gathered in community settings and two types of clinical settings: child and adolescent social care (CASC; Dutch: Jeugdgezondheidszorg [JGZ]), and child and adolescent mental healthcare (CAMH; Dutch: Jeugd Geestelijke Gezondheidszorg [Jeugd GGZ]). Additionally, in community settings data were gathered using the Child Behavior Checklist (Achenbach, 1991a), the Youth Self Report (Achenbach, 1991b) and the Intelligence and Development Scales (Grob, Hagmann-von Arx, Ruiter, Timmerman, & Visser, 2018). Table 1.1 provides an overview of the data available per setting. Community samples 1 and 2 and CAMH sample 1 were available for all studies in this thesis. Community sample 3, CASC sample 1 and CAMH sample 2 were only available for the studies presented in Chapters 5 and 6.

Table 1.1 Available SDQ, CBCL/YSR and IDS-2 data from the community, CASC, and CAMH settings

Setting	Sample	N	SDQ			CBCL/YSR			IDS-2
			Self and parent	Only self	Only parent	Self and parent	Only self	Only parent	
Community	1	519	274	217	28	276	211	26	
	2	443	206	220	17	192	181	1	220
	3	331	292	15	24				
CASC	1	124	31	74	19				
CAMH	1	4,053	3,493	206	354				
	2	229	177	39	13				

Notes. SDQ = Strengths and Difficulties Questionnaire; CBCL = Child Behavior Checklist; YSR = Youth Self Report; IDS–2 = Intelligence and Development Scales 2; CASC = Child and adolescent social care; CAMH = Child and adolescent mental health.

Outline. Construct validity aspects are investigated starting in *Chapter 2*. Chapter 2 provides an indication of whether the SDQ scales each measure a single and distinguishable domain of psychosocial behaviour by assessing the presumed five-factor structure of the self-report and parent-report SDQ versions in community and clinical settings. Additionally, the chapter provides information on the comparability of self-reported and parent-reported SDQ scores across these settings by investigating their measurement invariance. In this study we used SDQ data collected in community (samples 1 and 2) and CAMH (sample 1) settings.

The investigation into construct validity aspects continues in *Chapter 3*. Chapter 3 focuses on using the self-report and parent-report SDQ versions in a community setting. The chapter presents further information on the SDQ versions' presumed five-scale structures when used in community settings by conducting a more in depth assessment of their factor structures. Additionally, an investigation into associations between the SDQ scales and 1) conceptually similar CBCL/YSR scales, 2) conceptually different CBCL/YSR scales, and 3) conceptually different IDS-2 scales provides an indication of the extent to which each SDQ scale measures the domain of psychosocial behaviour it is presumed to measure (i.e., convergent and discriminant validity). In this part of Chapter 3, we used SDQ, CBCL/YSR and IDS-2 data of 962 adolescents, collected in community samples 1 and 2.

Criterion validity aspects are investigated starting at the end of *Chapter 3*. That part of the chapter provides indications of how well the total difficulties scale of both informant versions can be used to distinguish between adolescents from the general population and adolescents that are at risk of psychiatric disorders. Next, the chapter provides information of how well each of the five strengths and difficulties scales of both informant versions can be used to distinguish between adolescents from the general population and adolescents diagnosed with a disorder that content-wise matches the SDQ scale (Anxiety/Mood disorder for the emotional difficulties scale, CD/ODD for the conduct difficulties scale, ADHD and the hyperactivity/inattention scale, and ASD for the social problems and prosocial behaviour scales). For this investigation we used SDQ data collected in community (samples 1 and 2) and CAMH (sample 1) settings.

The investigation into criterion validity aspects continues in *Chapter 4*. Chapter 4 focuses on using the SDQ versions in a diagnostic context by examining how well diagnosed disorders (Anxiety/Mood disorder, CD/ODD, ADHD, and ASD) can each be predicted from separate SDQ scales of both informant versions. This examination provides information on how well SDQ scales can be used to provide a preliminary indication of the type of disorder an adolescent is suffering from. For this examination, SDQ data collected in the CAMH setting (sample 1) were used.

The examinations of criterion validity aspects presented in Chapters 3 and 4 are expanded upon in *Chapter 5*. Chapter 5 focuses on using SDQ score profiles that combine all self-reported and parent-reported SDQ scales, for distinguishing between adolescents from the community, CASC and CAMH settings, and for distinguishing between diagnosed disorders, including combinations of disorders. This investigation provides an indication of how useful the SDQ score profiles are for identifying individuals at risk of psychiatric disorders in a screening context and obtaining a preliminary indication of the type of disorder(s) in a diagnostic context. In this study we used SDQ data from the community, CASC and CAMH settings (all samples).

Chapter 6 presents joint community-based relative norms and gender-specific community-based relative norms per year of age, for use among Dutch adolescents aged 12 to 17. These norms are intended for interpreting adolescent self-reported and parent-

reported SDQ scale scores gathered in community and clinical settings. The norm scores were established using SDQ data collected in the community setting (all samples).

I conclude in *Chapter 7* by discussing the main findings from the studies presented in this thesis, describing the studies' main strengths and limitations, deriving implications for practice and providing recommendations for future research.

For practical and environmental reasons, the appendices to this thesis are made available online. Links to these appendices are provided in the chapters.

2

Psychometric properties of the Dutch Strengths and Difficulties Questionnaire (SDQ) in adolescent community and clinical populations

This chapter is based on:
Vugteveen, J., de Bildt, A., Serra, M., de Wolff, M. S., & Timmerman, M. E. (2018). Psychometric properties of the Dutch Strengths and Difficulties Questionnaire (SDQ) in adolescent community and clinical populations. *Assessment*.
https://doi.org/10.1177/1073191118804082

ABSTRACT

This study assessed the factor structures of the self-report and parent-report SDQ versions and their measurement invariance across settings based on clinical (n = 4,053) and community (n = 962) samples of Dutch adolescents aged 12 to 17. Per SDQ version, confirmatory factor analyses were performed to assess its factor structure in clinical and community settings and its measurement invariance across these settings. The results suggest measurement invariance of the presumed five-factor structure for the parent-report version and a six-factor structure for the self-report version. Further, evaluation of the SDQ scale sum scores as used in practice, indicated that working with sum scores yields a fairly reasonable approximation of working with the favourable but less easily computed factor scores. These findings suggest that self-reported and parent-reported SDQ scores can be interpreted using community-based norm scores, regardless of whether the adolescent has been referred for mental health problems or not.

INTRODUCTION

The Strengths and Difficulties Questionnaire (Goodman, 1997) aims at measuring psychosocial functioning among children and adolescents aged 4 to 17. This widely used questionnaire is valued for three reasons. Firstly, with only 25 items, the SDQ is relatively short. Secondly, the SDQ not only covers deficits (hyperactivity/inattention, conduct problems, emotional problems, peer problems), but also strengths (prosocial behaviour). Thirdly, the availability of multiple informant versions allows an individual´s psychosocial behaviour to be assessed from multiple perspectives. For adolescents aged 11 to 16, a self-report version and a parent-report version can be completed. A teacher version is also available, but as adolescents no longer spend the vast part of their school day with one or two teachers, teachers are increasingly often passed over as informants during adolescence.

The SDQ is typically used for screening and clinical assessment purposes. The usefulness of an instrument for these purposes can be judged against the standards of evidence-based assessment (Hunsley & Mash, 2007; Youngstrom & Frazier, 2013). According to these standards, an instrument is useful if it can be applied to predict an important criterion, prescribe a certain type of treatment or monitor an individual's progress (Youngstrom & Frazier, 2013). With these applications in mind, sound evidence for an instrument's psychometric properties is regarded as an essential prerequisite (Youngstrom, 2013). For the use of the SDQ among adolescents, multiple studies have provided insight into the psychometric properties of the self-report and parent-report SDQ versions (Goodman, 2001; van de Looij-Jansen, Goedhart, de Wilde, & Treffers, 2011; van Roy, Veenstra, & Clench-Aas, 2008). Two matters warrant further investigation. First, although the presumed five-factor structure (Goodman, 1997; Goodman, 2001) of both the self-report and parent-report SDQ versions has repeatedly been investigated in community settings, it has hardly been in clinical settings. Second, although the measurement invariance of both SDQ versions across demographic variables such as age, gender, and ethnicity has been investigated among adolescents, measurement invariance across adolescent community and clinical settings has not been addressed previously. The aim of the present study was to address these issues.

For the SDQ *parent-report* version, the few previous studies yielded support for the presumed five-factor structure of this SDQ version in community populations (He, Burstein, Schmitz, & Merikangas, 2013; van Roy et al., 2008) and a clinical population (Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004). However, the findings in the clinical population are of limited value for adolescents, since the clinical sample consisted of both adolescents and children without distinguishing between the two.

For the SDQ *self-report* version, the presumed five-factor structure has not been investigated in clinical populations. In community populations, several studies addressed

this matter. Some studies confirmed the five-factor structure (Goodman, 2001; Lundh, Wångby-Lundh, & Bjärehed, 2008; Richter, Sagatun, Heyerdahl, Oppedal, & Røysamb, 2011; Ruchkin, Koposov, & Schwab-Stone, 2007; van Roy et al., 2008), while others could only partially confirm it or could not (Bøe, Hysing, Skogen, & Breivik, 2016; Giannakopoulos et al., 2009; Koskelainen, Sourander, & Vauras, 2001; Ortuño-Sierra, Fonseca-Pedrero, Paino, Sastre i Riba, & Muñiz, 2015; Rønning, Handegaard, Sourander, & Mørch, 2004; van de Looij-Jansen et al., 2011). The mixed nature of the results can possibly be explained by differences in sample characteristics. For instance, all studies were performed among youths between the ages of 10 and 19, but some studies covered that whole age range while others only covered two or three years of age (e.g. 14-15 or 16-18). The samples further differed in country of origin; most of the studies mentioned were performed in North-West Europe, whereas others were performed in Greece, Russia, Spain and the United States. Cultural differences may underlie differences in the way the SDQ measures psychosocial functioning.

Considering the somewhat mixed results on the tenability of the five-factor structure regarding the SDQ self-report version, an alternative six-factor solution has been investigated (van Roy et al., 2008). This six-factor solution consists of the five factors as intended by Goodman (Goodman, 1997), and an additional *positive construal method* factor. The latter is comprised of the positively worded items, five in total, from the four difficulties scales. Such positively worded items tend to cluster together based on item stem similarity, regardless of the trait that they are supposed to measure (Pilotte & Gable, 1990; Schriesheim & Hill, 1981). The positive construal method factor thus expresses the method effect bias resulting from combining positively and negatively worded items in the SDQ difficulties scales.

Besides further investigation into how each SDQ version measures psychosocial functioning among adolescents in clinical and community settings, research is needed on whether the SDQ measures strengths and difficulties in the same way in both settings. The latter is highly relevant as it provides insight into the comparability of SDQ scores obtained in a clinical setting and SDQ scores obtained in a non-clinical setting. To sensibly compare SDQ scores across settings, measurement invariance is a prerequisite. A violation of measurement invariance occurs, for instance, when adolescents who complete the SDQ for the clinical assessment purposes at an institution for youth mental health care, interpret questions differently from adolescents who complete the questionnaire as part of a general health checkup at school. This would be problematic because it would mean that a very same SDQ score gathered in the two settings can bear a different meaning in terms of severity of the adolescents' problems. We are aware of only one study examining measurement invariance across community and clinical settings: Smits and colleagues (Smits, Theunissen, Reijneveld, Nauta, & Timmerman, 2016) found evidence for measurement invariance across these populations for the five-factor parent-report SDQ version among 2- to 14-year-olds. To the best of our knowledge, measurement invariance across these settings has not been investigated among adolescents.

The aim of the current study is to assess the presumed five-factor structure of the SDQ self-report and the parent-report versions, and to examine their measurement invariance across community and clinical populations of Dutch adolescents aged 12 to 17. In case the presumed five-factor structure does not fit adequately, we will investigate the fit of the six-factor structure, including the *positive construal method factor*. Additionally, this study assesses the way the SDQ scores are currently calculated in practice: summing item scores per SDQ scale, using equal weighting of items per scale. For the *parent-report* version we hypothesize to find confirmation for the presumed five-factor structure in the community and in the clinical populations, corroborating previous findings (Becker et al., 2004; He et al., 2013; van Roy et al., 2008). Further, we hypothesize to find measurement invariance of the five-factor SDQ parent-report version across the two populations, consistent with findings by Smits and colleagues (Smits et al., 2016), thereby assuming that the parent's manner of judgement regarding an adolescent's psychosocial functioning does not substantially differ from their manner of judgement of younger children's psychosocial functioning. As the five-factor structure closely resembles how SDQ scale scores are calculated in practice (i.e., summing item scores per scale), we hypothesize to find support for this sum score method.

For the SDQ *self-report* version, we cautiously expect to find confirmation for the presumed five-factor structure as findings from previous research regarding its factor structure in community populations are mixed. With regard to the factor structure of the self-report SDQ in a clinical population and this SDQ version's measurement invariance across community and clinical populations, we deem our study to be exploratory because these aspects were not covered by previous studies. Additionally, we do not have expectations of the extent to which our findings will support the sum score method as used in practice to calculate SDQ scale scores.

METHODS

Participants

Clinical sample. The clinical sample consists of 12- to 17-year old adolescents who, between January 1st of 2013 and December 31st 2015, were referred for the first time to one of 29 clinics of an institution for child and adolescent psychiatry in the North of the Netherlands. A total sample of 5,081 adolescents was eligible for this study. During the intake assessment, as part of routine outcome monitoring, data were collected online from these adolescents and their parents. For 4,053 of them, self-reported SDQ data (n = 354), parent-reported SDQ data (n = 206) or both (n = 3,493) were available. Among these adolescents the mean age was 14.2 years (SD = 1.6) among males (46.9%), and 14.6 years (SD = 1.5) among females (51.6%). Table 2.1 presents additional demographic and geographic characteristics of the clinical sample.

Table 2.2 provides an overview of the DSM-IV diagnoses, as established by trained professionals in a multidisciplinary team, generally consisting of at least a child- and adolescent psychiatrist and a child psychologist, supplemented with additional professionals such as a specialized nurse. Of the 4,053 adolescents in the sample, 2,812 had received a diagnosis in any of the four categories that content-wise respond to the SDQ scales. The remaining adolescents were not diagnosed with a DSM-IV disorder or their diagnosis was unknown (*n* = 628, 15,5%) or had received other DSM diagnoses (*n* = 613, 15.1%). The second column of the table shows that Anxiety/mood disorders were most prevalent, and Conduct/Oppositional Defiant Disorder (CD/ODD) were least prevalent. Per DSM-IV disorder (row), columns three through six provide information about the co-occurrence of disorders. Most prevalent is Attention-Deficit/Hyperactivity Disorder (ADHD) within the group with CD/ODD.

Table 2.1 Demographic and geographic characteristics of the adolescents in the clinical and community sample

	Clinical	Community
Characteristics	*N* **(%)**	*N* **(%)**
Gender		
Male	1,902 (46.9)[a]	474 (49.3)[b]
Female	2,093 (51.6)	482 (50.1)
Native country mother		
the Netherlands	[c]	754 (78.4)[d]
Other	[c]	149 (15.5)
Educational level mother		
Low	[c]	187 (19.4)[e]
Medium	[c]	281 (29.2)
High	[c]	282 (29.3)
Geographical region of the Netherlands		
North	2,563 (63.2)[f]	51 (5.3)[g]
East	1,452 (35.8)	164 (17.0)
South	4 (0.1)	155 (16.1)
West	24 (0.6)	367 (38.1)
Age		
12	581 (14.3)[h]	56 (5.8)
13	741 (18.3)	315 (32.7)
14	767 (18.9)	281 (29.2)
15	799 (19.7)	117 (12.2)
16	678 (16.7)	107 (11.1)
17	487 (12.0)	77 (8.0)

Notes. [a] Missing: *n* = 58 (1.4%); [b] Missing: *n* = 6 (0.6%); [c] information not available; [d] Missing: *n* = 100 (10.5%); [e] Missing: *n* = 212 (22.0%); [f] Missing: *n* = 10 (0.3%); [g] Missing: *n* = 225 (23.4%); [h] Missing: *n* = 9 (0.9%); [h] Missing: *n* = 9 (0.9%)

Table 2.2 Prevalence of DSM-IV diagnoses and comorbidity between DSM-IV diagnoses

		Co-occurring with ...			
DSM category[a]	***N*[b]**	**ADHD[c]**	**CD/ODD[c]**	**Anxiety/mood disorder[c]**	**ASD[c]**
ADHD	913	-	.18	.14	.16
Anxiety/Mood disorder	1,372	.09	.03	-	.09
ASD	719	.20	.04	.18	-
CD/ODD	391	.42	-	.09	.08

Notes. [a] ADHD: Attention-Deficit/Hyperactivity Disorder, ASD: Autism Spectrum Disorder, CD/ODD: Conduct/Oppositional Defiant Disorder; [b] The numbers in this column add up to more than 2,812 (number of adolescent in the sample with a diagnosis in any of the four categories) due to comorbidity; [c] The proportion of adolescents within each DSM category (row), also diagnosed with any of the other disorders

Community sample. Within the community sample of 12- to 17-year-old adolescents data were collected in three waves. The first wave of self-reported and parent-reported SDQ data were collected in 2009 and 2010, in the east, south and west of the Netherlands. The data were collected as part of a routine well-child care check provided regularly to all Dutch adolescents during their second year in secondary education (13- or 14-year-olds). The second wave of data, also collected among 13- or 14-year-old adolescents, consisted only of self-reported SDQ data and was collected in 2010 at six secondary schools in the west of the Netherlands. The sample resulting from these two waves consists of 519 adolescents for whom self-reported SDQ data (n = 217), parent-reported SDQ data (n = 28) or both (n = 274) were available. The third wave of data consisted of self-reported and parent-reported data and was gathered in 2016 and 2017 via schools throughout the Netherlands as part of a norming study of an intelligence test. The resulting sample consists of 443 adolescents for whom self-reported SDQ data (n = 220), parent-reported SDQ data (n = 17) or both (n = 206) were available.

In total, the community sample consisted of 962 adolescents, for whom self-reported SDQ data (n = 437), parent-reported SDQ data (n = 45) or both (n = 480) were available. Within this group the mean age was 14.1 years (SD = 1.4) among males (49.3%) and 14.2 years (SD = 1.4) among females (50.1%). Other demographic and geographic characteristics of the community sample are presented in Table 2.1. When compared to summary statistics published by Statistics Netherlands (2015), the community sample appears to be representative of the Dutch adolescent population regarding gender, ethnicity and mothers' educational level.

Table 2.1 presents information about the age distribution within the clinical and community samples. This information shows that 13- and 14-year-old adolescents are more heavily represented in the community sample (62.6%) than in the clinical sample (37.2%). This overrepresentation results from the initial data gathering as part of the well-child care check, which is provided to adolescents at approximately the age of 13 or 14.

Strengths and Difficulties Questionnaire

Adolescents and their parents completed the Dutch version of the self-report and parent-report SDQ versions, respectively (Van Widenfelt, Goedhart, Treffers, & Goodman, 2003). The 25-item questionnaires both consist of four subscales of five items focusing on difficulties relating to behaviour, emotional functioning, hyperactivity and interaction with peers, and one subscale of five items focusing on prosocial behaviour, which is considered a strength (Goodman, 1997). For each item, a three-point rating scale (0 = *not true*, 1 = *somewhat true* and 2 = *certainly true*) rates the degree to which the attribute is applicable to the adolescent. Five positively worded items belonging to different SDQ scales are reverse-coded. High scores on the four difficulties scales represent a high degree of difficulties; a high score on the prosocial behaviour scale represents a high degree of prosocial behaviour. As is recommended in the SDQ's scoring manual, SDQ scale scores were calculated by summing the item scores per scale while accounting for missing values as long as no more than two item scores per scale are missing. This method is called the sum score method in this paper.

Statistical analysis

Missing data. The clinical sample contained no missing data; the community sample data set contained some missing data at item level for the SDQ self-report version (M = 0.33%, SD = 0.32, min = 0%, max = 1.2%) and the SDQ parent-report version (M = 0.38%, SD = 0.28, min = 0%, max = 0.8%). Considering the small number of missing data we opted for two-way imputation with normally distributed errors to impute these data (van Ginkel, Ark, & Sijtsma, 2007).

Measurement invariance. First, the presumed five-factor structure, or in case the presumed five-factor does not fit adequately the six-factor structure, was modelled using single group (i.e., setting) confirmatory factor analysis (CFA) for ordinal data (Muthén, 1984).

This resulted in four single group CFA's, one for each setting (2: clinical, community) per SDQ version (2: adolescent, parent). Second, measurement invariance of the SDQ versions across settings was evaluated using multiple-group CFA models for ordinal data (Millsap & Yun-Tein, 2004). Per SDQ version, a set of four successive multiple-group CFA models (described below) was estimated. Each model within a set imposed additional constraints on the preceding model in order to examine whether the parameters of the models were equal across clinical and community settings, and thus whether measurement invariance would apply.

The first in each set of measurement invariance models was used to test configural invariance across settings. Configural invariance implies that the hypothesized factor structure (i.e., the position of the non-zero loadings) holds across both the clinical and community settings. For identification of the model, the following constraints were

applied (Millsap & Yun-Tein, 2004): In both settings, item intercepts were fixed to zero and the variances of the common factors to one; in the reference setting (i.e. the clinical setting), the residual variance of each continuous latent response variable was fixed to one and the mean of each common factor to zero; one threshold per variable and one additional threshold for the first item loading on each factor were constrained to be equal across settings.

If the configural invariance model fitted insufficiently, covariances between pairs of item residuals were allowed. To determine which covariance(s) to allow, we selected one residual covariance to free in the model using the modification indices of item pairs that belonged to the same factor, thereby selecting the one with the largest modification index among the indices with a value larger than ten, and the model was re-run. We repeated this process until the model fitted sufficiently or the model was re-run ten times. We chose ten residual covariances as the limit, because we considered allowing that many covariances or more to be an indication of factors beyond the factors tested. If the final five-factor model would not fit adequately, we fitted the six-factor model using the same procedure.

Next, measurement invariance models were estimated to test metric, strong and strict invariance, respectively. Metric invariance implies the equivalence of the factor loadings across settings. Strong invariance implies that SDQ factors and their underlying items are of equal meaning in both settings. Strict invariance implies that the latent trait was measured identically in both settings. Each consecutive model imposed additional constraints to its preceding model: equal factor loadings across settings (metric), equal thresholds across settings (strong), and equal residual variances across settings (strict).

All CFA models were estimated using Mplus version 8 (Muthén, & Muthén, 2017), using weighted least squares mean and variance adjusted (WLSMV) estimation. For illustration purposes, perturbed data and example code are available on https://osf.io/d5k7j/. The goodness-of-fit of the models was assessed by considering the root-mean-square error of approximation value (Steiger, 1980) and the comparative fit index (Bentler, 1990). We consider RMSEA values ≤. 08 combined with CFI values ≥ .90 to be acceptable, while RMSEA values ≤ .06 together with CFI values ≥ .95 are preferred, as is recommended by Hu and Bentler (Hu & Bentler, 1999). The goodness-of-fit of the measurement invariance models was additionally assessed by considering the change in CFI (ΔCFI), which represents the change in CFI value between pairs of successive models. Ideally model fit does not decrease from one model to the next. In other words, the CFI values should stay more or less the same. We considered a decrease of .01 or less as acceptable (Cheung & Rensvold, 2002). The fit measures mentioned take the number of model parameters into account. Consequently, fit statistics may indicate a more constrained model to fit slightly better than its preceding less constrained model purely as a result of the decreased number of parameters. For the sake of completeness and comparability with similar studies, Tucker-Lewis Index (Tucker & Lewis, 1973) values, chi-square values, their corresponding degrees

of freedom, and the chi-square Difftest outcomes are also presented. The TLI values were not interpreted, because they are highly correlated with the above mentioned CFI values and do not provide much additional information. Besides, the CFI is a more commonly used fit measure than the TLI. The Chi-square information was not interpreted, because the accuracy of chi-square tests relies heavily on the assumption that scores are normally distributed (Satorra, 1990) and thus often misrepresent the data.

Selecting a model per SDQ version. Per SDQ version, the presumed five-factor structure was evaluated first, because it most closely resembles how the SDQ is used in practice. The five-factor solution was selected for further examination if the RMSEA and CFI values showed sufficient fit. In case they did not, the fit of the six-factor alternative was evaluated with the same sequence of single group and multiple-group CFA's as described above.

For the selected model per SDQ version, effect size , indicating the number of standard deviations that the means of the clinical and community sample differ from each other, was used to interpret differences in factor means between the two settings (Choi, Fan, & Hancock, 2009). We considered effect sizes ≥ .50 as medium, and ≥ .80 as large.

The reliability per SDQ scale was estimated through the Omega coefficient (McDonald, 1999), which is a suitable measure as it allows unequal item loadings per factor (non-tau-equivalence) and allows residual item variances to be uncorrelated. SDQ scales are considered sufficiently reliable when Omega ≥ .70, while ≥.80 is preferred (Evers et al., 2010). Cronbach's alpha is reported for the sake of comparability to other studies.

Evaluating the sum score method as used in practice. In practice each SDQ scale score is calculated by summing the item scores of the items pertaining to that particular scale while accounting for missing values as long as no more than two item scores per scale are missing. The five-factor structure evaluated in this study resembles that method in the sense that it assumes the same division of items over factors. Unlike the sum score method, the five-factor structure does not assume equal weighting across items per factor, and takes dependency between factors into account. As a result, the factor scores associated with the five-factor CFA solution are not necessarily equal to the sum scores. Per SDQ version and SDQ scale, the use of the sum score method was evaluated by examining the association, expressed as Spearman rank correlation coefficients (rho), between the sum scores and the factor scores of the factor in the CFA associated with that SDQ scale. Note that the positive construal method factor from the six-factor model was not taken into account as no corresponding SDQ scale exists. We consider Spearman rho's >.85 to be supportive of the continued use of sum scores in practice.

RESULTS

The SDQ self-report version

Table 2.3 presents the goodness-of-fit statistics of the single group CFA's in the clinical and community settings. The table further presents the goodness-of-fit statistics for the successive multiple-group CFA models used to test measurement invariance across these settings.

Presumed five-factor model. The single group CFA's for the SDQ self-report version yielded acceptable RMSEA values and insufficient CFI values for both settings (clinical: RMSEA = .067, CFI = .850; community: RMSEA = .046; CFI = .896).

The configural invariance model, the first in the set of successive models to test measurement invariance, yielded acceptable RMSEA and insufficient CFI values (RMSEA = .062, CFI = .859, see configural invariance model I). Modification indices showed interpretable item residual covariances between multiple item pairs. Each item pair consisted of items belonging to the same factor. With ten of these residual item covariances allowed, model fit was still insufficient, with the RMSEA value being acceptable and the CFI value insufficient (RMSEA = .056, CFI = .892, see configural invariance model II). Consequently, the metric, strong and strict invariance models were not estimated.

Six-factor model. The single group models showed acceptable RMSEA and CFI values for the community setting, and acceptable RMSEA value but insufficient CFI value for the clinical setting (clinical: RMSEA = .061, CFI = .883; community: RMSEA = .034; CFI = .945).

The configural invariance model yielded an acceptable RMSEA value and an insufficient CFI value (RMSEA = .055, CFI = .894, see configural invariance model I). Allowing item residual covariances between one item pair resulted in acceptable model fit (RMSEA = .053, CFI = .902, see configural invariance model II). Acceptable fit was also found for the models measuring metric, strong and strict invariance (metric: RMSEA = .051, CFI = .904; strong: RMSEA = .050, CFI = .905; strict: RMSEA = .049, CFI = .904), indicating measurement invariance across settings. Figure A2.1 (available in the appendix on https://osf.io/d5k7j/) shows a representation of this model. The factor loadings, residual covariances, factor means and factor (co)variances of the strict invariance model are presented in Table 2.4.

Adolescents in the community and clinical settings differed from each other regarding their mean psychosocial strengths and difficulties scores: compared to the community setting, lower factor means were found in the clinical setting for the factors concerning difficulties (emotional difficulties: $\hat{d} = -1.63$; conduct problems: $\hat{d} = -1.08$; hyperactivity/attention problems: $\hat{d} = -1.49$; social problems: $\hat{d} = -0.97$), with the effect sizes being large. The settings did not significantly differ from each other with regard to the factor means for the strengths factor and the positive construal methods factor (prosocial behaviour: = 0.06, positive construal methods: $\hat{d} = -0.07$).

Table 2.3 Goodness-of-fit statistics of the presumed five-factor structure and the six-factor structure for the SDQ self-report version

Model	χ2	*df*	*p*-value	χ2 Difftest	*df* Difftest	*p*-value	RMSEA	RMSEA 90% CI	CFI	ΔCFI	TLI
Five-factor model as hypothesized by Goodman (Goodman, 1997)											
Single group											
Clinical	4,885.508	265	<.001				.067	[.066-.069]	.850		.831
Community	772.988	265	<.001				.046	[.042-.049]	.896.		.883
Multiple group											
Configural inv. I	5,451.699	530	<.001				.062	[.061-.064]	.859		.840
Configural inv. II[a]	4,271.369	510	<.001				.056	[.054-.057]	.892		.873
Six-factor model (including the positive construal method factor)											
Single group											
Clinical	3,862.007	255	<.001				.061	[.059-.062]	.883		.863
Community	525.249	255	<.001				.034	[.030-.038]	.945		.935
Multiple group											
Configural inv. I	4,210.048	510	<.001				.055	[.054-.057]	.894		.875
Configural inv. II[b]	4,593.298	518	<.001				.053	[.052-.055]	.902		.884
Metric fact. inv.	3,879.459	532	<.001	119.060	24	<.001	.051	[.050-.053]	.904	.002	.892
Strong fact. inv.	3,852.673	551	<.001	53.286	19	<.001	.050	[.049-.052]	.905	.001	.897
Strict fact. inv.	3,901.390	577	<.001	128.589	26	<.001	.049	[.048-.051]	.904	.001	.901

Notes. Configural inv. I = Configural invariance model with no freed item residual covariances; Configural inv. II = Configural invariance model with freed item residual covariances; Metric fact. Inv. = Metric factorial invariance model; Strong fact. Inv. = Strong factorial invariance model; Strict fact. Inv. = Strict factorial invariance model. Clinical group: n = 3,847; Community group: n = 917.

[a] Item residuals of ten item pairs (Q1 and Q4, Q1 and Q17, Q2 and Q10, Q2 and Q15, Q4 and Q17, Q9 and Q20, Q10 and Q15, Q15 and Q25, Q16 and Q24, Q18 and Q22) freed; [b] Item residuals of one item pairs (Q2 and Q10) freed

Table 2.4 Unstandardized parameter estimates and standard errors of the six-factor strict invariance model for the SDQ self-report version

SDQ scale	Item	SDQ scale factor loading	PCM factor loading	Threshold 1	Threshold 2
ES	Q3	0.63 (.02)		-0.26 (.02)	0.86 (.03)
	Q8	1.18 (.04)		-0.98 (.04)	0.52 (.03)
	Q13	1.59 (.06)		-0.29 (.04)	1.49 (.05)
	Q16	1.03 (.03)		-0.95 (.03)	0.46 (.03)
	Q24	1.20 (.04)		0.29 (.03)	1.72 (.04)
CP	Q5	1.02 (.05)		-0.26 (.03)	1.50 (.05)
	Q7	0.16 (.05)	0.81 (.06)	-0.77 (.03)	1.74 (.05)
	Q12	0.69 (.04)		0.94 (.03)	2.33 (.06)
	Q18	0.69 (.03)		0.19 (.02)	1.26 (.03)
	Q22	0.51 (.03)		1.15 (.03)	2.18 (.05)
HP	Q2	0.77 (.03)		-0.71 (.03)	0.77 (.03)
	Q10	0.84 (.04)		-0.59 (.03)	0.68 (.03)
	Q15	1.68 (.08)		-2.02 (.08)	0.15 (.04)
	Q21	0.46 (.04)	0.66 (.04)	-0.79 (.03)	1.41 (.04)
	Q25	1.07 (.04)	0.13 (.03)	-1.42 (.04)	0.88 (.03)
SP	Q6	0.79 (.04)		-0.24 (.03)	1.22 (.03)
	Q11	0.42 (.03)	0.12 (.03)	1.06 (.03)	1.65 (.03)
	Q14	0.84 (.04)	0.38 (.03)	0.48 (.03)	2.60 (.07)
	Q19	0.81 (.04)		0.81 (.03)	1.96 (.05)
	Q23	0.54 (.03)		0.05* (.02)	1.23 (.03)
PB	Q1	1.37 (.08)		-3.80 (0.15)	-0.77 (.04)
	Q4	0.63 (.03)		-1.85 (.04)	-0.41 (.02)
	Q9	0.82 (.04)		-2.23 (.05)	-0.51 (.02)
	Q17	0.81 (.04)		-2.79 (.08)	-1.11 (.04)
	Q20	0.69 (.03)		-1.41 (.03)	0.41 (.02)

Residual covariances

Q2-Q10	.42 (.02)

Factor means

	Clinical setting	Community setting	$\hat{d}$
ES	0	-0.97 (.05)	-1.63
CP	0	-1.50 (.10)	-1.08
HP	0	-0.91 (.05)	-1.49
SP	0	-0.85 (.07)	-0.97
PB	0	0.04* (.05)	0.06
PCM	0	-0.08* (.09)	-0.07

Factor (co)variances

	Clinical setting						Community setting					
	ES	CP	HP	SP	PB	PCM	ES	CP	HP	SP	PB	PCM
ES	1						0.75					
CP	0.21	1					0.37	1.80				
HP	0.31	0.56	1				0.31	0.68	0.89			
SP	0.62	0.26	0.13	1			0.57	0.75	0.20	1.23		
PB	0.03*	-0.54	-0.25	-0.22	1		-0.01*	-0.63	-0.22	-0.35	0.84	
PCM	-0.18	0.68	0.45	-0.14	-0.64	1	-0.09	0.43	0.32	-0.07*	-0.55	0.91

Notes. ES = emotional symptoms, CP = conduct problems, HP = hyperactivity/attention problems, SP = social problems, PB = prosocial behaviour, PCM = positive construal method *$p > .01$. For all other values $p < .01$.

Adequate reliability was found for the SDQ emotional difficulties, hyperactivity/inattention difficulties, and prosocial behaviour scales in the clinical and community settings, respectively (emotional difficulties: ω = .85, ω = .81; hyperactivity/inattention: ω =.80, ω = .79; prosocial behaviour: ω = .77, ω = .74). The conduct problems scale and the social problems scale showed to be insufficiently reliable in the clinical setting (conduct problems: ω = .65; social problems: ω = .69), and adequately reliable in the in the community setting (conduct problems: ω = .76, social problems: ω = .73).

The SDQ parent-report version

Table 2.5 presents the goodness-of-fit statistics of the single group CFA's in the clinical and community settings, and for the successive multiple-group CFA models used to test measurement invariance across these settings.

Presumed five-factor model. The single group models show insufficient RMSEA and CFI values for the clinical setting (RMSEA = .082, CFI = .848) and acceptable RMSEA and CFI values for the community setting (RSMEA = .048; CFI = .926).

The configural invariance model, yielded an acceptable RMSEA value and an insufficient CFI value (RMSEA = .075, CFI = .862, see configural invariance model I). The second configural invariance model, allowing item residual covariances for five item pairs, yielded acceptable RMSEA and CFI values (RMSEA: .064, CFI: .902, configural invariance model II). The metric invariance model yielded acceptable RMSEA and CFI values (RMSEA = .061, CFI = .907), as did the strong invariance model (RMSEA = .059, CFI = .909) and the strict invariance model (RMSEA: .058, CFI = .910). These results indicate measurement invariance across settings. Figure A2.2 (available in the appendix on https://osf.io/d5k7j/) shows a representation of the strict invariance model; the factor loadings, residual covariances, factor means and factor (co)variances are presented in Table 2.6.

Parental responses in the community and clinical settings differed from each other regarding their mean psychosocial strengths and difficulties scores, as can be seen in Table 2.6. Compared to the clinical setting, lower factor means for the factors concerning difficulties and a higher factor mean for the strengths factor were found in the community setting (emotional difficulties: $\hat{d}$ = –1.61; conduct problems: $\hat{d}$ = –1.19; hyperactivity/inattention problems: $\hat{d}$ = –1.41; social problems: $\hat{d}$ = –0.88, and prosocial behaviour: $\hat{d}$ = 0.65), with the effect sizes regarding the difficulties factors being large and the effect size for the strengths factor being medium.

Adequate reliabilities were found for all scales in the clinical and community setting, respectively (emotional difficulties: ω = .81, ω = .83; conduct problems: ω = .81, ω = .76; hyperactivity/inattention problems: ω = .80, ω = .83; social problems: ω = .77, ω = .82; prosocial behaviour: ω = .82, ω = .83).

Table 2.5 Goodness-of-fit statistics of the presumed five-factor structure for the SDQ parent-report version

Model	χ^2	*df*	*p*-value	χ^2 Difftest	*df* Difftest	*p*-value	RMSEA	RMSEA 90% CI	CFI	ΔCFI	TLI
Five-factor model as hypothesized by Goodman (Goodman, 1997)											
Single group											
Clinical	6,843.082	265	<.001				.082	[.080-.084]	.848		.828
Community	580.887	265	<.001				.048	[.042-.053]	.926		.916
Multiple group											
Configural inv. I	6,785.219	530	<.001				.075	[.073-.076]	.862		.844
Configural inv. II[a]	4,972.085	518	<.001				.064	[.062-.065]	.902		.887
Metric fact. inv.	4,759.011	538	<.001	62.924	20	<.001	.061	[.059-.063]	.907	.005	.896
Strong fact. inv.	4,660.638	558	<.001	74.201	20	<.001	.059	[.057-.061]	.909	.002	.903
Strict fact. inv.	4,661.278	589	<.001	199.904	31	<.001	.058	[.056-.059]	.910	.001	.907

Notes. Configural inv. I = Configural invariance model with no freed item residual covariances; Configural inv. II = Configural invariance model with freed item residual covariances; Metric fact. Inv. = Metric factorial invariance model; Strong fact. Inv. = Strong factorial invariance model; Strict fact. Inv. = Strict factorial invariance model. Clinical group: n = 3,699; Community group: n = 525.
[a] Item residuals of five item pairs (Q2 and Q10, Q8 and Q13, Q9 and Q20, Q15 and Q25, Q18 and Q22) freed

Table 2.6 Unstandardized parameter estimates and standard errors of the five-factor strict invariance model for the SDQ parent-report version

SDQ scale	Item	SDQ scale factor loading	Threshold 1	Threshold 2
ES	Q3	0.49 (.02)	-0.34 (.02)	0.54 (.02)
	Q8	0.93 (.04)	-1.17 (.04)	0.10 (.03)
	Q13	1.02 (.04)	-0.62 (.03)	0.90 (.03)
	Q16	1.22 (.05)	-1.25 (.04)	0.29 (.03)
	Q24	1.19 (.05)	0.07* (.03)	1.47 (.05)
CP	Q5	0.85 (.03)	-0.21 (.03)	1.04 (.03)
	Q7	1.23 (.05)	-0.50 (.03)	1.47 (.05)
	Q12	1.01 (.04)	1.12 (.04)	2.51 (.07)
	Q18	0.99 (.04)	0.09 (.03)	1.39 (.04)
	Q22	0.66 (.03)	0.92 (.03)	1.66 (.04)
HP	Q2	0.69 (.03)	-0.16 (.02)	0.97 (.03)
	Q10	0.61 (.03)	-0.08 (.02)	0.80 (.03)
	Q15	1.12 (.05)	-1.50 (.05)	-0.21 (.03)
	Q21	1.21 (.05)	-0.98 (.04)	0.80 (.04)
	Q25	0.98 (.04)	-1.17 (.04)	0.27 (.03)
SP	Q6	0.58 (.03)	-0.40 (.02)	0.67 (.03)
	Q11	0.82 (.04)	0.37 (.03)	1.40 (.04)
	Q14	1.56 (.09)	0.56 (.05)	3.07 (.13)
	Q19	0.88 (.04)	0.44 (.03)	1.67 (.04)
	Q23	0.55 (.03)	0.23 (.02)	1.26 (.03)
PB	Q1	2.84 (.33)	-3.91 (.40)	0.44 (.08)
	Q4	1.04 (.04)	-1.96 (.05)	-0.50 (.03)
	Q9	0.83 (.03)	-1.85 (.04)	-0.46 (.03)
	Q17	0.79 (.04)	-2.62 (.07)	-1.20 (.04)
	Q20	0.61 (.03)	-0.85 (.03)	0.50 (.02)

Residual covariances	
Q2-Q10	0.55 (.02)
Q8-Q13	0.55 (.02)
Q9-Q20	0.42 (.02)
Q15-Q25	0.51 (.02)
Q18-Q22	0.64 (.02)

Factor means			
	Clinical setting	Community setting	$\hat{d}$
ES	0	-1.69 (.08)	-1.61
CP	0	-1.21 (.08)	-1.19
HP	0	-1.33 (.07)	-1.41
SP	0	-1.09 (.09)	-0.88
PB	0	0.61 (.07)	0.65

Table 2.6 (continued)

Factor (co)variances										
	Clinical setting					Community setting				
	ES	CP	HP	SP	PB	ES	CP	HP	SP	PB
ES	1					1.16				
CP	0.13	1				0.43	0.70			
HP	0.10	0.73	1			0.53	0.63	1.27		
SP	0.47	0.41	0.25	1		0.89	0.43	0.53	1.49	
PB	-0.08	-0.71	-0.39	-0.50	1	-0.26	-0.44	-0.40	-0.73	1.04

Notes. ES = emotional symptoms, CP = conduct problems, HP = hyperactivity/attention problems, SP = social problems, PB = prosocial behaviour *$p > .01$. For all other values $p < .01$.

Evaluating the sum score method used in practice

Table 2.7 shows Spearman rank correlations between the SDQ scale sum scores, which resemble current practice, and factor scores resulting from the CFA analyses. All correlations provided support for the continued use of sum scores in practice, with correlations for the SDQ self-report version ranging from .90 for conduct problems scale to .98 for the hyperactivity/attention problems scale, and for SDQ parent-report version ranging from .92 for the prosocial behaviour scale to .97 for the emotional problems scale. For the sake of comparability with other studies, Table 2.7 additionally presents Cronbach's alpha coefficient per SDQ scale.

Table 2.7 Per SDQ version and scale, Cronbach's alpha and Spearman rank correlation coefficients between SDQ scale scores and factor scores

	SDQ self-report version		**SDQ parent-report version**	
SDQ scale	**Six-factor model**	**Cronbach's alpha**	**Five-factor model**	**Cronbach's alpha**
ES	.976	.79	.973	.78
CP	.900	.60	.933	.74
HP	.967	.77	.959	.78
SP	.908	.56	.925	.68
PB	.931	.64	.916	.75

Notes. ES = emotional symptoms, CP = conduct problems, HP = hyperactivity/attention problems, SP = social problems, PB = prosocial behaviour. For all correlation coefficients: $p < .01$.

DISCUSSION

This study evaluated the presumed five-factor structure and, if necessary, an alternative factor structure of the self-report and parent-report SDQ versions in clinical and community samples of Dutch adolescents aged 12 to 17. Next, measurement invariance of these factor structures across clinical and community settings was investigated. Finally, we evaluated the method of calculating SDQ scale scores as used in practice.

SDQ self-report version: Factor structure and measurement invariance. For the SDQ self-report version, the presumed five-factor structure was not supported, in both clinical and community settings. Our study was the first to assess the fit of the five-factor structure in a clinical setting, which prevents us from comparing our results to previous findings. With regard to the community setting our findings are in line with some previous studies (Koskelainen et al., 2001; van de Looij-Jansen et al., 2011), but not others (Ruchkin et al., 2007; van Roy et al., 2008). Neither differences in age range nor in cultural background seem to provide an explanation as our observations are in accordance with findings from some previous studies within samples with a similar age range (Giannakopoulos et al., 2009; Koskelainen et al., 2001; Rønning et al., 2004; van de Looij-Jansen et al., 2011) but not others (Ruchkin et al., 2007; van Roy et al., 2008), and our findings are in line with findings from some studies also performed in north-western European adolescent samples (Koskelainen et al., 2001; Rønning et al., 2004; van de Looij-Jansen et al., 2011) but not all (van Roy et al., 2008).

For the SDQ self-report version, the alternative six-factor solution was preferred over the five-factor solution, suggesting that the presence of reverse-worded items in the difficulties scales affects the SDQ's factor structure. The six-factor structure was found to fit the community data acceptably well, as is in line with findings from Van Roy and colleagues (van Roy et al., 2008). Regarding the clinical data, this factor structure was not fully confirmed to fit adequately. Model fit for both settings improved to an acceptable level by allowing item residuals of one pair of items to covary. Allowing this covariance accounts for the presence of a minor factor within one of the factors, as will be explained in more detail later. Further, evidence was found for measurement invariance of this six-factor structure across clinical and community settings. This finding suggests that the SDQ self-report version is useful for screening purposes, as this SDQ version measures adolescents' strengths and difficulties in the same way in clinical (e.g., during intake preceding thorough diagnostic assessment by clinicians) and community settings (e.g., as part of a routine well-child check-up or at school).

SDQ parent-report version: Factor structure and measurement invariance. For the SDQ parent-report version, the five-factor structure was supported for the community setting, which is in line with previous findings in similar samples (He et al., 2013; van Roy et al., 2008). Regarding the clinical data, we could not fully confirm the fit of this factor structure. Allowing some item residuals to covary improved model fit in both settings. Further, evidence was found for measurement invariance of the five-factor structure across clinical and community settings, as was hypothesized. Extending upon Smits and colleagues' (Smits et al., 2016) similar observations regarding children, our findings suggest that the SDQ parent-report version measures adolescents' strengths and difficulties in the same way in clinical and community settings.

Allowing item residual covariances. From the CFA's we learned that some item pairs contributed to their factor and additionally had something else in common, which called for allowing the item residuals of these items to covary. One of these item pairs, items 2

('restless, overactive') and 10 ('constantly fidgeting or squirming') of the hyperactivity/inattention problems factor, was found for both SDQ versions (i.e., the five-factor model for the SDQ parent-report version and the six-factor model for the SDQ self-report version). This finding is consistent with findings from several previous studies among adolescents (Bøe et al., 2016; Ortuño-Sierra et al., 2015; Rønning et al., 2004; Smits et al., 2016; van de Looij-Jansen et al., 2011; van Roy et al., 2008). Within the same factor, items 15 ('easily distracted, concentration wanders') and 25 ('sees tasks through to the end') seemed to have something other than belonging to the same factor in common for the SDQ parent-report version. This finding too is in accordance with findings from a number of previous studies (Bøe et al., 2016; Ortuño-Sierra et al., 2015; Smits et al., 2016). The persistent findings regarding these two item pairs most likely indicate the presence of minor factors hyperactivity and/or inattention within the hyperactivity/inattention factor (Bøe et al., 2016; van de Looij-Jansen et al., 2011). This is not surprising as the hyperactivity/inattention factor's name already suggests heterogeneity within the factor. Although the need for allowing some item residuals to covary indicates that the items measuring the two constructs can to some extent be distinguished from each other, the CFA results imply that the items within the hyperactivity/inattention factor are strongly associated, and together can be used to sensibly measure hyperactivity/inattention.

Scale reliabilities per SDQ version. As was described above, both SDQ versions were found to be measurement invariant, and thus can be used to distinguish at risk adolescents from others *across* settings. Additionally, the scale reliabilities can be used to assess how useful the scales of both SDQ versions are for the purpose of differentiating between adolescents *within* each setting. With the exception of the conduct and social difficulties scales of the SDQ self-report version in the clinical setting, all SDQ scales of both SDQ versions were found to be sufficiently reliable in both settings. For the conduct and social difficulties scales, the clinical setting data show limited variance in scores compared to the community setting data, resulting in lower reliabilities.

Evaluating SDQ scales as currently used in practice. Apart from evaluating the factor structure, the aim of our study was to assess the way the SDQ scores are currently calculated in practice: summing item scores per SDQ scale, using equal weighting of items per scale. This summing method was supported for both SDQ versions by the findings of the current study, as SDQ scale sum scores and its associated factor scores were all highly correlated. This indicated that although unequal weighting of items per SDQ scale would be optimal, the currently used equal weighting yields a fairly reasonable approximation. For the SDQ self-report version, evidence was found for a six-factor structure including a positive construal method factor. Methodologically this factor is interesting, because it indicates an unintended effect of the positive wording of some items measuring difficulties. For practice, this methodological factor is less interesting as it does not contribute to measurement of psychosocial functioning content-wise.

Strengths and limitations

This study focused primarily on evaluating the presumed five-factor structure of the SDQ. If needed, an alternative factor structure was evaluated. It cannot be ruled out that a factor structure other than the ones under investigation would yield an even better representation. However, finding the best fitting factor structure was not the purpose of our study. Our aim was to evaluate factor structures that closely resemble how the SDQ is used in practice.

Our study is the first to assess measurement invariance of the self-report and parent-report SDQ versions across clinical and community settings. Knowledge about potential measurement invariance helps determine whether SDQ scores from clinical and community settings can be interpreted in the same way, and thus can be compared. Comparing scores across these settings is, for instance, important for clinicians as they are often interested in how a referred adolescent's scores compared to adolescents from a non-clinical population.

Further, the current study evaluated the factor structure and measurement invariance of multiple SDQ versions, whereas most other studies investigated the psychometric properties of only one informant version. During adolescence, adolescents themselves are increasingly often used as the informant, but self-reports are potentially more prone to social desirability and biased estimation of their own psychosocial functioning than reports from other informants are. Therefore, the parent is also a frequently used informant. From investigating both versions within similar adolescent samples, we, for instance, learned that reverse-worded items affect the factor structure of the SDQ self-report version. For the parent-report version, measurement invariance was found without having to take into account the reverse-worded nature of some of the items.

The current study is subject to four potential limitations. First, approximately half of community sample data were collected about seven years before the rest of the data were collected. By handling these data as if it were one community sample, we assume that adolescents' and parents' interpretation of the items and thus the factor structure of both SDQ versions has not changed over time. We consider this assumption tenable, given the relatively short time span of about seven years between collecting both parts of the sample. The tenability of this assumption is further supported by the fact that we found measurement invariance across settings.

The second limitation of the current study is that clinical and community samples are not comparable based on geographical origin and age distribution. The adolescents in the community sample mainly reside in the west, south and east of the Netherlands, while the adolescents in the clinical sample mainly reside in the north and east of the Netherlands. In the worst case scenario, we may have assessed measurement invariance across geographic regions instead of across settings. The Netherlands is a small and relatively densely populated country, which are characteristics that likely reduce the interpretational differences across geographic regions. Therefore, we deem it to be fairly improbable that our findings regarding measurement invariance are biased by these sample differences. With respect to age, the two samples are incomparable as 13- and 14-year-old adolescents

are overrepresented in the community sample. As both samples further contain substantial numbers of 12- and 15- to 17-year-olds and the total age range of our sample is relatively small, we have no reason to believe that this sample difference would cause a violation of measurement invariance of either SDQ version under investigation in this study.

Third, we have not been able to compare the clinical and community samples on characteristics as migration background and social economic status as we had no indicators of these characteristics for the adolescents in the clinical sample and indirect indicators of these characteristics for the community sample. These factors may have confounded our findings.

Fourth, if necessary we adapted our models by using modification indices to determine which, if any, residuals variances to allow, as is a commonly used approach in similar studies. This course of action results in models that are to some extent sample dependent, which may have biased our results. Therefore, we hope that others will try to replicate our findings in other but similar samples.

Implications

The SDQ is used in clinical and community settings, albeit for different purposes. In community settings, mainly consisting of adolescents that do not suffer from psychosocial problems, SDQ scores are used to screen for adolescents at risk of developing psychiatric disorders. In clinical settings, mainly consisting of adolescents with psychosocial problems, SDQ scores are often used to provide a preliminary indication of the problems at hand, which is then more thoroughly considered by clinicians. Although the aim of the use of the SDQ differs across settings, our findings indicate measurement invariance across settings, meaning that the SDQ screens for psychosocial problems in the same way in both settings.

In practice, the SDQ is used to assess an adolescent's psychosocial functioning by comparing the adolescent's SDQ scale scores to community-based norm scores. The scale scores are calculated by summing the item scores per scale. This method is insightful and easy to work with, but also quite blunt as it assumes that all items within a scale measure the construct equally well. For the five scales of both SDQ versions strong association were found between sum scores and factor scores, which can be regarded as support for the continued use of the sum score method in practice. Note that the positive construal method factor in the six-factor structure for the self-report version was not evaluated for use in practice, because this is a methodological factor that does not contribute to measurement of psychosocial functioning content-wise. These findings are encouraging for clinical and community practice as they suggest that SDQ scores of adolescents can be interpreted using community-based norm scores, regardless of whether the adolescent has been referred for mental health problems or not.

Our findings further show the conduct and social difficulties scales of the SDQ self-report version to be insufficiently reliable within the clinical setting. This suggests that these scales are of limited use for the purpose of differentiating between adolescents within a clinical setting.

3

Validity aspects of the self-report and parent-report Strengths and Difficulties Questionnaire (SDQ) versions among Dutch adolescents

This chapter is based on:
Vugteveen, J., de Bildt, A., Theunissen, M., Reijneveld, S.A., & Timmerman, M. (2019). Validity Aspects of the Strengths and Difficulties Questionnaire (SDQ) Adolescent Self-Report and Parent-Report Versions Among Dutch Adolescents. *Assessment*. https://doi.org/10.1177/1073191119858416

ABSTRACT

In this study validity aspects of the Strengths and Difficulties Questionnaire (SDQ) self-report and parent-report versions were assessed among Dutch adolescents aged 12 to 17 years (community sample: n = 962, clinical sample: n = 4,053). The findings mostly support the continued use of both SDQ versions in screening for psychosocial problems, as a) exploratory structural equation analyses partially supported the grouping of items into five scales, b) investigation of associations between scales of the SDQ and the Child Behavior Checklist, Youth Self Report and Intelligence Development Scales 2 provided evidence for the SDQ versions' convergent and divergent validity, and c) receiver operating characteristics (ROC) curves yielded evidence for both SDQ versions' criterion validity by showing that these questionnaires can be used to screen for psychosocial problems in general, except for the self-report version for males. Regardless of the adolescent's gender, the ROC curves showed both SDQ versions to be useful for screening for three specific types of problems: Anxiety/Mood disorder, Conduct/Oppositional Defiant Disorder, and Attention-Deficit/Hyperactivity Disorder. Additionally, parent-reported SDQ scores can be used to screen for Autism Spectrum Disorder.

INTRODUCTION

Psychosocial problems frequently occur in adolescents, with the prevalence estimated at 15 to 25% (Fergusson et al., 1993; Ormel et al., 2015). To screen for these problems in community settings, for example during large scale general health check-ups, the Strengths and Difficulties Questionnaire (Goodman, 1997; Goodman, 1999) is a widely used instrument. The SDQ is particularly suitable for this purpose as it a) is relatively short, b) focuses on strengths (prosocial behaviour) as well as multiple types of difficulties (emotional problems, conduct problems, hyperactivity/inattention, peer problems), and c) is available in multiple informant versions (self-report, parent, teacher). Of the informant versions, the teacher version is least likely to be relevant for use among adolescents, because adolescents spend only a limited amount of time with each of their teachers. To be of use for screening purposes in an adolescent community population, the SDQ should be of good validity for this population. As relatively few studies examined the SDQ's validity among adolescents, the purpose of this study was to examine a broad range of validity aspects of the SDQ self-report and parent-report versions among Dutch adolescents. That is, we considered evidence for their presumed internal structure, and their convergent, discriminant, and criterion validity.

Internal structure. The SDQ was designed to measure strengths as well as four types of difficulties, resulting in a presumed five-factor structure. For the SDQ *self-report* version, this five-factor structure showed to be tenable in some studies among adolescents (Goodman, 2001; Lundh et al., 2008; Richter et al., 2011; Ruchkin et al., 2007; van Roy et al., 2008), but not in others (Bøe et al., 2016; Giannakopoulos et al., 2009; Koskelainen et al., 2001; Ortuño-Sierra et al., 2015; Rønning et al., 2004; van de Looij-Jansen et al., 2011). It is important to note that none of the studies mentioned can be compared directly to the others, because they strongly differ concerning, for instance, sample age range and country of origin. Another study found a six-factor solution to fit, rather than the presumed five-factor solution (van Roy et al., 2008). This six-factor structure includes the presumed five factors and an additional *positive construal method* factor. The additional factor consists of the positively worded items, five in total, from the four difficulties scales, implying that this factor expresses the positive wording effects for items measuring difficulties. Note that the positive construal method factor in this six-factor model differs from the positive construal method factor in the modified five-factor model assessed by Van de Looij-Jansen et al. (2011). In their model, the prosocial behaviour factor was modified by adding cross-loadings onto the five positively worded items measuring difficulties. By doing so they ignored that, besides their positive wording, the items measuring prosocial behaviour are presumed to have in common that they measure strengths. The resulting factor thus represents a combination of a wording effect and prosocial behaviour, implying it is not just a wording factor. For the SDQ *parent-report*

version, the few studies that were conducted found support for the presumed five-factor structure (He et al., 2013; van Roy et al., 2008).

Convergent and discriminant validity. In previous studies, the SDQ's convergent validity has been investigated using the empirically based syndrome scales of the parent-reported Child Behavior Checklist (Achenbach, 1991a) and its self-report version, the Youth Self Report (Achenbach, 1991b), as gold standards. Like the SDQ, the CBCL and YSR belong to the domain of instruments measuring behaviour, and their validity is well documented (Achenbach, 1991a; Achenbach, 1991b; Chen, Faraone, Biederman, & Tsuang, 1994; Nakamura, Ebesutani, Bernstein, & Chorpita, 2009; van Lang, Ferdinand, Oldehinkel, Ormel, & Verhulst, 2005).

Concerning the SDQ's convergent validity, only a few studies were conducted among populations consisting of only adolescents. For the *SDQ self-report version*, moderate to strong correlations between conceptually similar SDQ and YSR scales were found (Van Widenfelt et al., 2003; Vogels, Siebelink, Theunissen, de Wolff, & Reijneveld, 2011). For the *SDQ parent-report version*, the only study among adolescents we found, showed moderate correlations between conceptually similar scales of the two instruments (Vogels et al., 2011). Note that the above mentioned studies differed in which of the eleven CBCL/YSR empirically based syndrome scale(s) they regarded as conceptually similar to each SDQ scale. One of the studies compared all SDQ scales to only the three broadband CBCL/YSR scales (i.e., externalizing problems: delinquent and aggressive behaviour; internalizing problems: anxious/depressed, somatic complaints, withdrawn; total problems: sum of all problem items; Vogels et al., 2011), thereby generating only generic results. The two other studies additionally considered the eight specific CBCL/YSR scales (e.g., aggressive behaviour, anxious/depressed) by linking each SDQ scale to one or more (Van Widenfelt et al., 2003) syndrome scales.

Of the studies mentioned above, only Van Widenfelt and colleagues (Van Widenfelt et al., 2003) considered an aspect of discriminant validity. They did so by reporting correlations between conceptually unrelated SDQ and CBCL/YSR syndrome scales. However, whether the convergent correlations (i.e., correlations between scores on related scales) were stronger than the discriminant correlations (i.e., correlations between scores on unrelated scales) was not tested. Note that all scales within a domain can be expected to be associated to some extent, because of the shared domain; conceptually related SDQ and CBCL/YSR scales can be expected to be strongly associated, whereas associations among conceptually unrelated SDQ and CBCL/YSR scales are expected to be weak.

We were not able to find studies that address the SDQ's discriminant validity by looking at associations between SDQ scales and scales from instruments belonging to unrelated domains, such as the domain of intelligence. Comparing scales across domains is useful because valid measurements of these different domains are expected to show weak or negligible associations.

Criterion validity. In the few studies we found among adolescent clinical and community samples, the SDQ's ability to distinguish between these two types of samples was found to be good for both the *SDQ self-report version* (Goodman et al., 1998; Vogels et al., 2011) and the *SDQ parent-report version* (Vogels et al., 2011).

Addressing the issues mentioned above, the aim of our study is to examine the internal structure and the convergent, discriminant and criterion validity of the SDQ self-report and parent-report versions among 12- to 17 year old Dutch adolescents, when used for screening purposes. First, we will assess both SDQ versions' factor structures among the community sample of adolescents, because we aim to evaluate the SDQ as it is used in screening. This screening setting resembles the context in which the data were collected, i.e. in a community setting. Note that in a previous study using the same data, the SDQ's measurement invariance across clinical and community populations was supported (Vugteveen, de Bildt, Serra, de Wolff, & Timmerman, 2018), which assures us that we do not unintentionally ignore a potential setting effect by looking at only the community data. Here, first we will assess the presumed five-factor structure of both SDQ versions using confirmatory factor analysis (CFA), because this structure most closely resembles how SDQ scale scores are calculated in practice. In case the five-factor structure shows insufficient fit, the fit of a six-factor structure containing the presumed five factors and a *positive construal methods factor* will be evaluated. These two structures express that the items are perfect indicators of a single (or two) construct(s). As this rarely holds for psychological scales (Asparouhov & Muthén, 2009), we supplement the CFA results with a more exploratory approach: exploratory structural equation modelling (Asparouhov & Muthén, 2009). As far as we know, ESEM has only been used on self-reported SDQ scores in one adolescent sample (Garrido et al., 2018), which yielded some support for the presumed five-factor structure, but also indicated items to contribute to scales other than their presumed scale. As further ESEM-based evidence is lacking, we are unsure of whether the presumed five-factor structure will be supported or not in our study.

Second, the SDQ versions' convergent and discriminant validity will be tested by investigating associations between the SDQ scales and conceptually similar CBCL/YSR scales (same domain), conceptually different CBCL/YSR (same domain), and conceptually different Intelligence and Development Scales (IDS-2; Grob, Hagmann-von Arx, Ruiter, Timmerman, & Visser, 2018). Considering the results from previous research, we expect to find evidence supporting the SDQ versions' convergent and divergent validity.

Third, we will assess the SDQ scales' ability to distinguish clinical groups from a community group, therewith focusing on the use of the SDQ in a screening context. This clearly differs from an earlier analysis of the clinical data used in this study, where the data were used to investigate how well SDQ scale scores of adolescents referred to mental health care can be used to predict specific types of disorders in a clinical context (Vugteveen et al., 2018). Here, we expect to find support for the use of both SDQ versions' total difficulties scale for distinguishing between the two general groups (community,

clinical). Further, as no substantial research is available on how well each of the five SDQ difficulties and strengths scales can be used to distinguish clinical groups with specific types of disorders from the community group, we have no hypotheses on this matter and we regard our investigation to be exploratory.

METHODS

Participants

Community sample. The community sample data of 12- to 17-year-old Dutch adolescents were collected in two waves. The first wave of data was collected in 2009/2010 at secondary schools, if possible as part of a routine well-child care check which is provided to all Dutch adolescents during their second year in secondary education (13- or 14-year-olds). For the 519 adolescents from this wave, adolescent self-reported data (n = 217), parent-reported data (n = 28), or both (n = 274) were available. Also available were YSR data (n = 211), CBCL data (n = 26), or both (n = 276). The second wave of data was gathered in 2016 and 2017 as part of a norming study of an intelligence test, resulting in adolescent self-reported SDQ data (n = 220), parent-reported SDQ data (n = 17), or both (n = 206) from 443 adolescents. Further, YSR data (n = 181), CBCL data (n = 1), or both (n = 192) were available for these adolescents. Additionally, IDS-2 data (n = 220) were gathered. Combining data from the two waves resulted in a community sample consisting of 962 adolescents, for whom adolescent-reported SDQ data (n = 437), parent-reported SDQ data (n = 45) or both (n = 480) were available. Also available for the adolescents in this sample were YSR data (n = 392), CBCL data (n = 27), or both (n = 468), and IDS-2 data (n = 220). Table A3.1 (appendices, indicated by A, are available on https://osf.io/dmjns/) provides an overview of the available questionnaires within the community sample. The mean age in this sample was 14.1 years (SD = 1.4) among males (49.6%) and 14.2 years (SD = 1.3) among females (50.4%).

Clinical sample. The 12- to 17-year-old adolescents in the clinical sample were referred for the first time to one of the clinics of an institution for child and adolescent psychiatry in the North of the Netherlands, between January 1st of 2013 and December 31st 2015. Their data were collected online during the intake assessment as part of routine outcome monitoring. Of the 4,053 adolescents in the clinical sample, 2,812 had received a DSM-IV diagnosis in any of the four categories that content-wise respond to the SDQ scales. Table A3.2 (available on https://osf.io/dmjns/) provides an overview of these diagnoses and an indication of co-occurrence of disorders within the sample. The diagnoses were established by trained professionals in a multidisciplinary team, generally consisting of at least a child- and adolescent psychiatrist and a child psychologist, and, depending on the context, supplementary professionals such as a specialized nurse. Within this sample,

adolescent-reported SDQ data (n = 354), parent-reported SDQ data (n = 206), or both (n = 3,493) were available. The mean age was 14.2 years (SD = 1.6) among males (47.6%), and 14.6 years (SD = 1.5) among females (52.4%).

Additional demographic and geographic characteristics of both samples are presented in Table 3.1. For comparison, summary statistics of the Dutch population are presented in the last column of the table (Statistics Netherlands, 2015).

Measures

The Strengths and Difficulties Questionnaire. The 25-item Dutch versions of the self-report and parent-report SDQ versions (Van Widenfelt et al., 2003) both consist of four five-item scales focusing on difficulties relating to emotional functioning, conduct, hyperactivity/inattention, and interaction with peers. These four scales together form the total difficulties scale. Additionally, the SDQ contains a five-item scale focusing on strengths in the form of prosocial behaviour (Goodman, 1997) The items are rated on a three-point rating scale (0 = *not true*, 1 = *somewhat true* and 2 = *certainly true*). Five positively worded items belonging to different SDQ difficulties scales are reverse-coded. High scores on the four difficulties scales, represent a high degree of difficulties; a high score on the prosocial behaviour scale represents a high degree of prosocial behaviour.

The Child Behavior Checklist and Youth Self-Report. The Dutch versions of the CBCL and YSR contain 113 and 112 items, respectively (Verhulst, Van der Ende, & Koot, 1996; Verhulst, Van der Ende, & Koot, 1997). The items are rated on a three-point rating scale (0 = *not true*, 1 = *somewhat or sometimes true* and 2 = *very true or often true*) (Achenbach, 1991a; Achenbach, 1991b). For both instruments, all but 17 (CBCL) or 10 items (YSR) can be divided into 8 empirically based syndrome scales with item numbers varying from 8 to 17 (YSR) or 18 (CBCL): 1) aggressive behavior, 2) anxious/depressed, 3) attention problems, 4) delinquent behavior, 5) somatic complaints, 6) social problems, 7) thought problems, 8) withdrawn. Five of these scales can be summarized in two broader scales: 1) the delinquent behavior and aggressive behavior scales form the externalizing behavior scale and 2) the withdrawn, somatic complaints and anxious/depressed scales are combined in the internalizing behavior scale. Together all items, including the items not belonging to the empirically based syndrome scales, form the total behavior problems scale. A second way to summarize 55 of the CBCL and 53 of the YSR items is by dividing them into six DSM-oriented scales: 1) affective problems, 2) anxiety problems, 3) attention/deficit/hyperactivity problems, 4) conduct problems, 5) oppositional defiant problems, and (6) somatic problems (Achenbach, 2014).

Table 3.1 Demographic and geographic characteristics of the adolescents in the clinical (n = 4,053) and community (n = 962) samples

	Clinical sample	Community sample	Dutch population
Characteristics	***N*** **(%[a])**	***N*** **(%[a])**	**%**
Gender			
Male	1,902 (47.6)[b]	474 (49.6)[c]	49.5
Female	2,093 (52.4)	482 (50.4)	50.5
Age			
12	581 (14.3)	56 (5.9)[d]	16.5
13	741 (18.3)	315 (33.1)	16.3
14	767 (18.9)	281 (29.5)	16.4
15	799 (19.7)	117 (12.3)	16.9
16	678 (16.7)	107 (11.2)	16.9
17	487 (12.0)	77 (8.1)	17.1
Mother's country of birth			
the Netherlands	[e]	754 (83.2)[f]	78.6
Other	[e]	149 (16.5)	21.4
Mother's educational level			
Low	[e]	187 (24.9)[g]	23.6
Medium	[e]	281 (37.5)	41.7
High	[e]	282 (37.6)	34.7
Geographical region of the Netherlands			
North	2,565 (63.4)[h]	51 (6.9)[i]	10.2
East	1,452 (35.9)	164 (22.2)	21.1
South	4 (0.1)	155 (20.9)	21.4
West	24 (0.6)	367 (49.9)	47.3

Notes. [a] Percentages computed of valid cases only. [b] Missing: n = 58; [c] Missing: n = 6; [d] Missing: n = 9; [e] information not available; [f] Missing: n = 100; [g] Missing: n = 212; [h] Missing: n = 10; [i] Missing: n = 222

The Intelligence and Development Scales. The Dutch version of the IDS-2 (Grob, Hagmann-von Arx, Ruiter, Timmerman, & Visser, 2018) contains measures of general intelligence and of five developmental domains. General intelligence is measured with fourteen subtests aimed at visual processing, long term memory, processing speed, short term memory (auditory), short term memory (spatial-visual), abstract thinking, and verbal thinking. The five developmental domains are measured with between two and four subtests per domain, including dividing attention (domain: executive functioning), visual motor skills (domain: psychomotor skills), recognizing emotions (domain: socioemotional competences), logical-mathematical thinking (domain: school skills), and conscientiousness (domain: motivation). All scales are normed, with the general intelligence scale expressed as IQ-scores (i.e., mu = 100, sigma = 15) and the five developmental domains as standardized scores (i.e. mu = 10, sigma = 3).

Statistical analysis

Missing data. Our data set contained missing data at two levels: questionnaire level and item level. First, for some participants entire SDQ, CBCL, YSR or IDS-2 questionnaires were unavailable resulting in missing data at questionnaire level. The sample description of both samples contains information about the available questionnaires. Second, the community sample data set contained some missing data at item level for the SDQ self-report version (M = 0.33%, SD = 0.32, min = 0.0%, max = 1.2%) and the SDQ parent-report version (M = 0.38%, SD = 0.28, min = 0.0%, max = 0.8%). This sample data set further contained some missing data at item level for the YSR within the group of adolescents that also filled in the SDQ (M = 0.69%, SD = 0.50, min = 0.1%, max = 4.4%); and for the CBCL within the group of parents that filled in the SDQ (M = 0.85%, SD = 0. 53, min = 0.2%, max = 4.2%). The missing data at questionnaire level was not imputed; analyses were performed based on available cases. Taking into account the small number of missing values at item level and the type of analyses we were planning to perform, these missing data were imputed in two ways. First, for the calculation of SDQ, YSR and CBCL scale scores, mean imputation of item scores was used, in compliance with the instruments' manuals. For the CBCL and the YSR, five parents and four adolescents had too many scores missing to calculate a score for the DSM oriented somatic problems scale; these item scores were not imputed, resulting in missing scale scores. All other missing item scores were imputed and scale scores were calculated. The resulting scale scores were used for analyses at scale level based on available cases: calculating mean scale scores and correlations between scale scores. Second, for analyses at item level, a single two-way imputation with normally distributed errors was used to impute the missing data (van Ginkel et al., 2007); this approach, unlike mean imputation, leads to unbiased item covariance estimates, which is preferred for item level analyses. The two-way imputed data were used for confirmatory factor analyses on the SDQ data and estimating the reliability of the SDQ, CBCL and YSR scales.

Among the adolescents in the community sample that had IDS-2 data available, some IDS-2 data were missing at domain level (M = 4.32%, SD = 3.48, min = 0.0%, max = 10.0%). Underlying are missing data at subtest level. We deemed it unwise to impute entire subtests and decided to perform the analyses regarding the IDS-2 data based on available cases.

Factor structure. The factor structures of the SDQ versions (adolescent, parent) were evaluated using the community sample data. Per SDQ version, the presumed five-factor structure was modelled using CFA for ordinal data (Muthén, 1984). The CFA models were estimated using weighted least squares mean and variance adjusted (WLSMV) estimation. Goodness-of-fit was assessed by considering the comparative fit index (Bentler, 1990) and the root mean square error of approximation value (Steiger, 1980). We consider CFI values ≥ .90 combined with RMSEA values ≤. 08 to be acceptable, while preferring CFI

values ≥ .95 combined with RMSEA values ≤ .06 (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). For comparability with other studies, Tucker-Lewis Index (Tucker & Lewis, 1973) values were also presented. In case the RSMEA and CFI values indicated insufficient fit of the five-factor model, the six-factor alternative was evaluated. This factor structure consists of the presumed five factors and an additional *positive construal method* factor containing five positively worded items from the four difficulties scales. The positively worded items of the prosocial behaviour scale were not included in this additional factor as these items differ from the five positively worded items measuring difficulties. They differ from each other in the sense that the prosocial items indicate a strength and jointly make up a single scale that does not contain any negatively worded items, whereas the positively worded items from the positive construal method factor are part of difficulties scales that contain both positively and negatively worded items.

One of the main characteristics of CFA is that it allows items to only load on the factor(s) they are presumed to contribute to, and it fixes other cross-loadings at zero. In our five-factor model this implies that each item has a freely estimated loading on a single factor only. In our six-factor model this implies that five items have freely estimated loadings on their presumed factor and on the positive construal method factor, all other items each have a freely estimated loading on a single factor only. Although this closely resembles how SDQ scale scores are calculated in practice, it may distort model fit (Marsh, Morin, Parker, & Kaur, 2014) and inflate associations between factors, which in turn affects the estimated factor loadings and factor reliabilities (Asparouhov, Muthén, & Morin, 2015). To overcome these limitations, we supplemented our analyses with exploratory structural equation models (ESEM) using WLSMV estimation and target rotation (Asparouhov & Muthén, 2009; Marsh et al., 2014). The latter aims to minimize cross-loadings without forcing them to be zero. As with CFA, we used ESEM to test the fit of the presumed five-factor structure. In case that model did not fit, we evaluated the fit of the six-factor structure. For all factor analyses, loadings ≥ .30 are regarded as salient loadings.

For CFA and ESEM models that showed sufficient fit, local fit was assessed using the standardized expected parameter change statistic (Saris, Satorra, & Van der Veld, 2009). SEPC values >.20 warranted allowing item residuals to correlate by freeing them one at the time, starting with the parameter associated with the largest SEPC, until acceptable local fit was found.

Scale reliabilities. Per SDQ scale, the reliability of the observed scores was computed using the nonlinear structural equation modelling reliability coefficient (Yang & Green, 2015), based on a one-factor model including correlated item residuals as far as necessary to achieve acceptable local fit. The reliability coefficient takes into account both the SDQ items' ordinal nature and allows for unequal item loadings per factor (non-tau-equivalence). SDQ scales were considered sufficiently reliable when $\rho_{NL} \geq .70$, while ≥ .80 was preferred (Evers et al., 2010). For the purpose of comparability with other studies,

Cronbach's alpha coefficients were calculated for all SDQ, CBCL and YSR scales. For the IDS-2, we lacked the item scores necessary to compute Cronbach's Alpha.

The analyses mentioned so far are analyses performed at item level. For the remaining analyses, scale level data were used.

Descriptive statistics. To characterize differences across informants and settings, mean scale scores were calculated per SDQ, CBCL and YSR scale. Note that SDQ scores were available for both settings (community, clinical), and all other instruments for the community setting only. In contrast to SDQ, CBCL and YSR scores, IDS-2 scores were normed, allowing us to compare community scores to population means. For this purpose, z-tests were used. To assess potential setting differences in SDQ scale scores per SDQ version, we conducted a multivariate analysis of variance (manova) with the SDQ scales as dependent variables and the setting as independent variable, followed by t-tests for post-hoc univariate comparisons per SDQ version and scale to compare scale scores across settings. Given the nature of the populations, it is to be expected that the prevalence of psychiatric disorders related to psychosocial functioning was higher in the clinical sample than in the community sample. Therefore, we expect to find higher mean scale scores for the SDQ difficulties scales and a lower mean scale score for the SDQ strength scale in the clinical sample than in the community sample.

Convergent and discriminant validity. To express the strength of associations of rank scores on SDQ (adolescent, parent) and YSR (adolescent)/CBCL (parent) scale pairs, we computed Spearman Rho correlations. These correlations were computed for conceptually related SDQ and YSR/CBCL scale pairs, denoted as convergent correlations, and for conceptually different SDQ and CBCL/YSR or IDS-2 scale pairs, denoted as discriminant correlations. Per SDQ scale, Steiger's test (Steiger, 1980) was used to compare convergent correlations with discriminant correlations within the set of 1) eight empirically based syndrome scales, 2) eight empirically based syndrome scales and the three broader empirically based syndrome scales, and 3) six DSM-oriented scales.

Criterion validity. In order to determine how well both SDQ versions were able to distinguish between the community and clinical populations, we used receiver operating characteristic (ROC) curves. First, we investigated how well the total difficulties scale of both SDQ versions was able to distinguish between the two populations. Next, we examined each SDQ strengths and difficulties scale's ability to differentiate between the community population and a clinical subpopulation that had received a diagnosis content-wise corresponding to the particular SDQ scale (Anxiety/Mood disorder for the SDQ emotional difficulties scale, Conduct / Oppositional Defiant Disorder (CD/ODD) for the SDQ conduct difficulties scale, Attention-Deficit/Hyperactivity Disorder (ADHD) for the SDQ hyperactivity/inattention difficulties scale and Autism Spectrum Disorder (ASD)

for the SDQ social difficulties and prosocial behaviour scales). Additionally, we provided an investigation into potential gender differences. Area under the curve (AUC) values were reported as an index of discriminative ability. We considered AUC values ≥ .80 as indicating sufficient ability to distinguish between samples. For comparing AUC values of different SDQ scales, DeLong's test for paired ROC curves was used (DeLong, DeLong, & Clarke-Pearson, 1988).

For all statistical tests, a significance level of $\alpha = .01$ was used. The confirmatory factor and ESEM analyses were performed in Mplus version 8.0 (Muthén & Muthén, 2017). All other analyses were performed in R, version 3.4.1. (R Core Team, 2016). Data imputation was performed using the mokken package (Van der Ark, 2007), the ROC analyses were performed using the pROC package (Robin et al., 2011), and the ρ_{NL} coefficients were computed using the semTools package (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018). For illustration purposes, perturbed data and example code are available on https://osf.io/dmjns/.

RESULTS

Factor structure of SDQ self-report and parent-report versions

Table 3.2 presents the goodness-of-fit statistics of the CFA and ESEM models evaluated using community sample data. For the *self-report version*, the CFA models showed insufficient fit for the five-factor model and acceptable fit for the six-factor model, suggesting the potential presence of a wording effect. As both CFA models still may have misrepresented the SDQ's factor structure, the five-factor ESEM model was evaluated. This model showed excellent fit. Table 3.3 presents factor loadings and factor correlations for both CFA models and the ESEM model. Note that two items in the ESEM model (items 7 "obedient" and 11 "friend", both positively worded items measuring difficulties) showed negligible loadings on their intended factor (loadings ≤ .30) and one item (item 1 "considerate", prosocial factor) loaded on its intended factor as well as on the conduct difficulties factor.

Information about the local fit of the six-factor CFA model and the five-factor ESEM model is provided in Tables A3.3 (available on https://osf.io/dmjns/). Per model, three error correlations were added to the model, indicating that three item pairs formed subfactors within the factor they belong to. The estimated models are provided in Table A3.4 (available on https://osf.io/dmjns/) One additional item (item 5 "temper", conduct factor) now showed substantial loadings on its intended factor as well as on the emotional difficulties factor.

Table 3.2 Goodness-of-fit statistics of the CFA and ESEM models for the self-report and parent-report SDQ versions in the community sample

Model	χ^2	df	*p*-value	RMSEA	RMSEA 90% CI	CFI	TLI
SDQ self-report version							
CFA-5F	772.988	265	<.001	.046	[.042 - .049]	.896	.883
CFA-6F	525.249	255	<.001	.034	[.030 - .038]	.945	.935
ESEM-5F	304.576	185	<.001	.027	[.021 - .032]	.976	.960
SDQ parent-report version							
CFA-5F	576.368	265	<.001	.047	[.042 - .053]	.926	.916
ESEM-5F	274.950	185	<.001	.030	[.023 - .038]	.979	.965

Notes. ESEM = exploratory structural equation modelling, CFA = confirmatory factor analysis; for the SDQ self-report version: $n = 917$; for the SDQ parent-report version: $n = 525$. χ^2: chi square value; df: degrees of freedom; RMSEA: root mean square error of approximation; CFI: comparative fit index; TLI: Tucker-Lewis index; 5F: 5 factors; 6F: 6 factors

For the *parent-report version*, the five-factor CFA model fitted acceptably; the five-factor ESEM model fitted better. Table 3.4 presents factor loadings and factor correlations for both CFA models and the ESEM model. The ESEM model showed one item (item 5 "temper", conduct factor) loading negligibly on its intended factor (loading ≤ .30). This item and five other items (items 10 "fidgety", 14 "generally liked", 17 "kind", 19 "bullied", and 24 "fears") showed salient but weak loadings (loadings ranging from .30 to .37) on a factor they were not intended to load on.

For this SDQ version, information about the local fit of the five-factor CFA and ESEM models is provided in Tables A3.3 (available on https://osf.io/dmjns/). Four error correlations were added to the CFA model, and two were added to the ESEM model, indicating the presence of subfactors. The estimated models are provided in Table A3.5 (available on https://osf.io/dmjns/). One additional item (item 12 "temper", conduct factor) now showed a negligible loading on its intended factor.

Scale reliability

For the SDQ *self-report* version, ρ_{NL} estimates of .73, .55, .72, .56, and .63 were found for the emotional difficulties, conduct difficulties, hyperactivity/inattention problems, social problems and prosocial behaviour scales, respectively. Regarding the SDQ *parent-report* version, ρ_{NL} estimates for these scales were .71, .57, .72, .68, and .75. The estimates suggested questionable reliability for four out of five adolescent-reported SDQ scales and two out of five parent-reported SDQ scales. Cronbach's alpha coefficients per scale of both SDQ versions and the CBCL/YSR, are presented in Tables 3.5 and 3.6, respectively.

Table 3.3 Standardized parameter estimates of the CFA and ESEM models for the SDQ self-report version

	CFA five-factor model					CFA six-factor model						ESEM five-factor model				
Item/ factor	**ES**	**CP**	**HP**	**SP**	**PB**	**ES**	**CP**	**HP**	**SP**	**PB**	**PCM**	**ES**	**CP**	**HP**	**SP**	**PB**
3	**.49**					**.49**						**.54**	.21	.003	-.17	.04
8	**.72**					**.72**						**.67**	.10	0.02	.07	.17
13	**.79**					**.79**						**.75**	.12	-0.01	.05	.09
16	**.64**					**.64**						**.60**	-.18	.06	.12	-.08
24	**.78**					**.78**						**.72**	-.22	.06	.18	-.06
5		**.72**					**.78**					.25	**.49**	.16	.13	.02
7		**.45**					**.05**				**.36**	-.04	**.23**	.21	-.17	-.30
12		**.59**					**.61**					-.08	**.66**	.05	.03	-.06
18		**.64**					**.67**					-.13	**.55**	.16	.28	.01
22		**.60**					**.62**					-.09	**.53**	.02	.07	-.01
2			**.77**					**.79**				-.16	-.004	**.90**	.13	.13
10			**.73**					**.75**				.002	.01	**.75**	.13	.15
15			**.77**					**.79**				.15	.03	**.73**	-.13	-.002
21			**.57**					**.35**			**.40**	.05	.21	**.38**	-.14	-.19
25			**.64**					**.50**			**.28**	.12	-.01	**.55**	-.20	-.25
6				**.56**					**.64**			.15	-.11	-.03	**.60**	-.14
11				**.51**					**.40**		**.61**	.14	.24	-.09	**.15**	-.27
14				**.71**					**.58**		**.30**	.04	.18	.04	**.45**	-.25
19				**.68**					**.74**			.11	.19	.07	**.57**	.08
23				**.49**					**.55**			.14	.09	-.08	**.45**	-.01
1					**.77**					**.77**		.15	-.42	.03	-.07	**.52**
4					**.46**					**.45**		.01	.01	.03	-.22	**.42**
9					**.62**					**.62**		.06	.18	-.07	-.15	**.72**
17					**.64**					**.63**		-.02	-.13	-.05	-.07	**.49**
20					**.53**					**.54**		-.02	.06	-.02	.10	**.68**
Factor correlations																
	ES	CP	HP	SP	PB	ES	CP	HP	SP	PB	PCM	ES	CP	HP	SP	PB
ES	1.00	0.28	0.34	0.58	-0.02	1.00	0.33	0.37	0.59	-0.02	-.06	1.00	0.10	0.30	0.41	-0.03
CP		1.00	0.63	0.54	-0.62		1.00	0.52	0.50	-0.52	-.42		1.00	0.38	0.24	-0.34
HP			1.00	0.24	-0.31			1.00	0.17	-0.20	.35			1.00	0.11	-0.23
SP				1.00	-0.45				1.00	-0.28	-.09				1.00	-0.12
PB					1.00					1.00	-.71					1.00
PCM											1.00					

Notes. ESEM = exploratory structural equation modelling, CFA = confirmatory factor analysis, ES = emotional symptoms, CP = conduct problems, HP = hyperactivity/inattention problems, SP = social problems, PB = prosocial behaviour, PCM = positive construal method. Per item, its loading on its intended factor is printed in bold

Table 3.4 Standardized parameter estimates of the CFA and ESEM models for the SDQ parent-report version

	CFA five-factor model					ESEM five-factor model				
Item/ factor	**ES**	**CP**	**HP**	**SP**	**PB**	**ES**	**CP**	**HP**	**SP**	**PB**
3	**.34**					**.45**	.04	.05	-.19	.02
8	**.84**					**.85**	.04	-.02	.14	.13
13	**.79**					**.78**	-.14	.06	.14	.05
16	**.78**					**.56**	.26	-.01	.21	.04
24	**.78**					**.55**	.35	-.10	.26	.02
5		**.62**				**.34**	**.17**	.22	-.06	-.10
7		**.57**				**.06**	**.36**	.17	-.16	-.34
12		**.42**				**.08**	**.39**	.10	.15	.14
18		**.71**				**.10**	**.53**	.25	-.12	-.18
22		**.49**				**.16**	**.66**	-.05	-.08	-.02
2			**.78**			**-.25**	.16	**.80**	.24	.14
10			**.77**			**-.18**	.17	**.74**	.34	.24
15			**.86**			**.12**	-.05	**.84**	-.07	.01
21			**.61**			**-.01**	.17	**.50**	-.13	-.21
25			**.83**			**.18**	-.18	**.86**	-.20	-.14
6				**.53**		**.19**	-.09	-.09	**.41**	-.26
11				**.63**		**.03**	-.04	.08	**.59**	-.20
14				**.75**		**.18**	-.06	.13	**.40**	-.37
19				**.80**		**.33**	.04	.25	**.48**	.05
23				**.66**		**.17**	-.06	.04	**.58**	-.14
1					**.87**	**.01**	-.23	-.14	-.13	**.65**
4					**.78**	**-.08**	-.13	.14	-.23	**.67**
9					**.75**	**.15**	-.05	.02	-.19	**.77**
17					**.62**	**-.01**	.31	-.03	-.22	**.65**
20					**.61**	**.15**	-.16	.01	.14	**.77**
Factor correlations										
	ES	CP	HP	SP	PB	ES	CP	HP	SP	PB
ES	1.00	0.51	0.39	0.68	-0.21	1.00	0.19	0.35	0.36	-0.19
CP		1.00	0.67	0.46	-0.54		1.00	0.37	0.17	-0.12
HP			1.00	0.38	-0.28			1.00	0.17	-0.19
SP				1.00	-0.57				1.00	-0.22
PB					1.00					1.00

Notes. ESEM = exploratory structural equation modelling, CFA = confirmatory factor analysis, ES = emotional symptoms, CP = conduct problems, HP = hyperactivity/inattention problems, SP = social problems, PB = prosocial behaviour, PCM = positive construal method. Per item, its loading on its intended factor is printed in bold.

Table 3.5 Per SDQ version (self-report, parent-report) and per setting (community, clinical): Mean scale scores, standard deviations and Cronbach's Alpha

		SDQ version			
		Self-report[a]		Parent-report[b]	
Setting	SDQ scale	α[c]	*M* (*SD*)	α[c]	*M* (*SD*)
Community	Total[c]	.66	8.1 (4.8)	.70	6.4 (5.0)
	Emotional	.68	2.1 (2.0)	.69	1.6 (1.9)
	Conduct	.51	1.3 (1.3)	.46	0.8 (1.2)
	Hyper	.74	3.4 (2.3)	.78	2.4 (2.4)
	Social	.54	1.3 (1.5)	.64	1.5 (1.8)
	Prosocial	.61	8.0 (1.7)	.72	8.3 (1.8)
Clinical	Total	.70	14.5 (5.9)	.67	15.9 (6.5)
	Emotional	.77	4.4 (2.8)	.75	5.0 (2.8)
	Conduct	.58	2.5 (1.8)	.73	2.8 (2.4)
	Hyper	.76	5.3 (2.6)	.76	5.2 (2.8)
	Social	.54	2.3 (1.9)	.66	2.9 (2.3)
	Prosocial	.64	7.9 (1.8)	.74	7.4 (2.2)

Notes. SDQ: Strengths and Difficulties questionnaire; α: Cronbach's index of internal consistency (alpha); [a] Self-report version clinical setting: $N = 3,847$; community setting: $N = 3,699$; [b] Parent-report version clinical setting: $N = 917$; community setting: $N = 525$; [c] Per SDQ version, all mean scale score comparisons across settings, except the comparison for the adolescent-reported prosocial behaviour scale, indicated a significant difference with $p < .001$

Scale scores

Community setting mean scale scores of both SDQ versions and the CBCL/YSR are presented in Tables 3.5 and 3.6, respectively. Note that it is impossible to gain insight into relative problem levels in our sample by comparing mean scale scores within an instrument to each other, because some types of behaviour are generally less prevalent than the others. Table 3.7 provides community setting mean scale scores for the IDS-2. The IDS-2 scales were normed, allowing us compare our sample means to population means. Table 3.7 presents the outcomes of the z-tests that were used. The community sample scored significantly lower than the population on the general intelligence scale, but not on the five developmental domains.

Table 3.5 additionally presents mean scale scores for both SDQ versions in the clinical setting. The manova and post-hoc t-tests performed to assess potential setting differences in SDQ scale scores per SDQ version, showed significant setting-effects on all SDQ scales, except the adolescent-reported prosocial behaviour scale (t (4762) = 8.26, p = .16), with higher scores on the difficulties scales of both SDQ versions, and lower scores on the parent-reported SDQ prosocial behaviour scale, in the clinical setting than in the community setting (F (3,962) = 120.09, $p < .001$).

Convergent and discriminant validity

Table 3.8 presents Spearman rho correlations between the SDQ scales of the SDQ parent-report version and the CBCL (parent-reported) scales, and between the SDQ self-report version and the YSR (adolescent-reported) scales.

Table 3.6 For the adolescent self-reported YSR and the parent-reported CBCL: Mean scale scores, standard deviations and Cronbach's Alpha (community setting)

		Informant			
		Self-report (*N* = 850)		Parent-report (*N* = 489)	
YSR/CBCL scale		**α**	***M (SD)***	**α**	***M (SD)***
Empirically based syndrome scales	Aggressive problems	.81	3.7 (3.8)	.85	2.4 (3.4)
	Anxious/depressed	.84	3.5 (3.9)	.80	2.1 (2.8)
	Attention problems	.76	4.4 (3.1)	.81	3.0 (3.1)
	Delinquent	.69	3.1 (2.8)	.69	1.2 (1.9)
	Social problems	.69	2.7 (2.6)	.77	1.4 (2.3)
	Somatic complaints	.75	2.6 (2.8)	.63	1.5 (1.9)
	Thought problems	.72	2.7 (3.0)	.63	1.4 (2.0)
	Withdrawn	.73	2.6 (2.5)	.77	1.8 (2.3)
	Total	.93	23.4 (15.7)	.93	13.8 (12.5)
	Externalizing	.86	6.8 (6.0)	.87	3.6 (4.8)
	Internalizing	.89	8.8 (7.8)	.86	5.4 (5.6)
DSM-oriented scales	Affective problems	.78	3.3 (3.5)	.72	1.6 (2.3)
	Anxiety problems	.66	2.0 (2.0)	.66	1.0 (1.5)
	Attention problems	.76	4.2 (2.9)	.81	2.3 (2.6)
	Conduct problems	.71	2.5 (2.7)	.71	0.9 (1.7)
	Oppositional defiant problems	.63	1.6 (1.6)	.76	1.2 (1.6)
	Somatic problems*	.68	1.6 (2.0)	.54	1.1 (1.4)

Notes. YSR: Youth Self Report; CBCL: Child Behavior Checklist; α: Cronbach's index of internal consistentcy (alpha)
* YSR: *n* = 846 (scale score missing for 4 cases); CBCL: *n* = 484 (scale score missing for 5 cases)

Table 3.7 IDS-2 mean scale scores (community setting)

IDS-2	N	*M (SD)*
General intelligence	216	93.8 (16.9)[a]
Executive functioning	214	9.9 (2.2)[b]
Psychomotor skills	207	10.5 (2.1)[b]
Socioemotional competences	209	10.3 (3.1)[b]
School skills	215	9.5 (2.7)[b]
Motivation	198	10.4 (3.0)[b]

Notes. IDS-2: Intelligence Development Scale 2
[a] Significantly different from the normed population means (general intelligence: z = -6.07, p <.001, 99% CI [91.17, 96.43])
[b] Not significantly different from the normed population means (executive functioning: z = -0.49, p = .626, 99% CI [9.37, 10.43]; psychomotor skills: z = 2.40, p =.017, 99% CI [9.96, 11.04]; socioemotional competences z = 1.45, p = .148, 99% CI [9.77, 10.84]; school skills: z = -2.44, p = .015, 99% CI [8.97, 10.03]; Motivation: z = 1.88, p = .061, 99% CI [9.85, 10.95])

Convergent correlations (correlations between conceptually similar scales) are printed in bold; the remaining correlations are discriminant correlations (correlations between conceptually different scales). All but five of the resulting correlations were significantly different from zero, with convergent correlations ranging from .39 to .79 and discriminant correlations from .12 to .68. Per SDQ scale and for all but 13 comparisons, the convergent correlations were positive and significantly stronger than the discriminant correlations, in line with our expectations.

Table 3.9 presents Spearman rho correlations between the scales of both SDQ versions and the IDS-2 scales. Of the resulting correlations, which are all considered discriminant correlations, only 16 were significantly different from zero. These 16 correlations, ranging from -.38 to -.19, indicated the presence of weak negative relationships between SDQ and IDS-2 scores, which is in line with our expectations. All but four of these correlations were found between scales of the SDQ *self-report* version and IDS-2 scales, suggesting that adolescent self-reported SDQ scale scores were slightly more, but at most weakly, associated with the adolescent's intelligence than parent-reported scores.

Table 3.8 Spearman Rho correlations between SDQ scores and YSR/BCL scale scores (community setting)

YSR/CBCL scales Total		Scales SDQ self-report version*					Scales SDQ parent-report version*				
		Emotion	Conduct	Hyper	Social	Total	Emotion	Conduct	Hyper	Social	
Empirically based syndrome scales	Aggressive problems	.55	.33	**.45**	.45	.20	.57	.35	**.59**	.44	.24
	Anxious/depressed	.53	**.68**	.13	.25	.27	.42	**.56**	.22	.14	.25
	Attention problems	.65	.34	.35[a]	**.72**	.15	.68	.33	.40[a]	**.74**	.23
	Delinquent	.45	.20	**.43**	.37	.20	.46	.25	**.48**	.35	.22
	Social problems	.56	.47	.24	.33	**.39**	.58	.44	.36	.36	**.43**
	Somatic complaints	.47	.51	.18	.29	.17	.29	.45[a]	.14	.11**	.15
	Thought problems	.56	.45	.28	.40	.29[a]	.44	.36	.26	.34	.21
	Withdrawn	.53	.55	.16	.22	**.47**	.51	.41	.21	.21	**.54**
	Externalizing	.57	.31	**.49**	.47	.22	.59	.36	**.60**	.45	.25
	Internalizing	.62	**.71**	.18	.31	.36[b]	.54	**.61**	.25	.21	.42[b]
	Total	**.74**	.58	.40[b]	.55	.32	**.73**	.54[b]	.50[b]	.54	.36
DSM-oriented scales	Affective problems	.60	.56[c]	.26	.38	.34	.51	.45[c]	.31	.30	.33
	Anxiety problems	.51	**.62**	.12	.26	.26	.44	**.53**	.22	.22	.23
	Attention problems	.58	.24	.35[c]	**.74**	.05**	.67	.30	.41[c]	**.79**	.16
	Conduct problems	.44	.19	**.42**	.37	.17	.45	.23	**.52**	.36	.18
	Oppositional defiant problems	.45	.25	**.43**	.36	.16	.50	.28	**.55**	.39	.19
	Somatic problems†	.38	.43	.14	.23	.12	.23	.41	.11**	.08**	.08**

Notes. SDQ: Strengths and Difficulties Questionnaire; YSR: Youth Self Report; CBCL: Child Behavior Checklist; Correlations between conceptually similar scales (convergent correlations) are presented in bold.
Unlike the other discriminant correlations, this discriminant correlation is not significantly stronger than the lowest of the convergent correlations between the associated SDQ scale and each of the [a] eight empirically based CBCL/YSR scales, [b] all empirically based CBCL/YSR scales or [c] the DSM-oriented CBCL/YSR scales
* SDQ self-report version – YSR combination: n = 840; SDQ parent-report version – CBCL combination: n = 456
** Correlation not significant at the 0.01 level; all other correlations are significant at the 0.01 level
† YSR: *n* = 836 (4 cases missing); CBCL: *n* = 451 (5 cases missing)

Table 3.9 Spearman Rho correlations between SDQ scores and IDS-2 scale scores (community setting)

IDS-2 scales	Scales SDQ self-report version						Scales SDQ parent-report version					
	N	Total	Emotional	Conduct	Hyper	Social	N	Total	Emotional	Conduct	Hyper	Social
General intelligence	204	-.20*	.01	-.31*	-.01	-.33*	137	-.32*	-.15	-.21	-.19	-.30*
Developmental domains												
Executive functioning	202	-.15	.00	-.23*	.00	-.26*	136	-.21	-.12	-.06	-.10	-.27*
Motivation for school	187	-.28*	-.10	-.18	-.38*	.01	127	-.10	.07	-.14	-.19	-.01
Psychomotor skills	195	-.17	-.11	-.10	-.12	-.09	131	-.18	-.13	-.05	-.16	-.08
School skills	203	-.20*	-.07	-.24*	-.03	-.29*	136	-.24*	-.17	-.12	-.14	-.22
Socioemotional competences	197	-.19*	.06	-.28*	-.14	-.19*	134	-.08	-.10	-.08	-.04	-.20

Notes. SDQ: Strengths and Difficulties Questionnaire; IDS: Intelligence Development Scales.
* Correlation significant at the 0.01 level

Criterion validity

The AUC values presented in Table 3.10 indicated sufficient discriminative ability of all SDQ scales, except for the adolescent-reported social problems scale and the adolescent- and parent-reported prosocial behaviour scales. The latter were not corroborated as being insufficiently capable of distinguishing between the community sample and the clinical subsample of adolescents with an ASD diagnosis. It is noteworthy that for both SDQ versions, the emotional difficulties, the conduct problems and hyperactivity/inattention scales were better at distinguishing between types of disorders than the SDQ total difficulties scale was at distinguishing between the total community and clinical samples. Figures A3.1 to A3.10 (available on https://osf.io/dmjns/) show the ROC graphs associated with these results. Table A3.6, Table A3.7 and figures A3.11 to A3.30 (available on https://osf.io/dmjns/) provide an investigation of potential gender effects. The main gender difference was found for the SDQ self-report version's total difficulties scale, which distinguished sufficiently between the community and clinical samples for females (AUC = .84) but not for males (AUC = .76).

Table 3.10 Per SDQ version and scale, its ability to distinguish between community and clinical (sub)samples

	Self-report SDQ version			**Parent-report SDQ version**		
SDQ scale	***Comm. N***	***Clin. N[a]***	***AUC (SE)***	***Comm. N***	***Clin. N***	***AUC (SE)***
Total	917	3,847	.80 (.01)	525	3,699	.87 (.01)
Emotional	917	1,325	.87 (.01)	525	1,215	.92 (.01)
Conduct	917	363	.85 (.01)	525	346	.93 (.01)
Hyper	917	873	.85 (.01)	525	856	.91 (.01)
Social	917	667	.75 (.01)	525	670	.84 (.01)
Prosocial	917	667	.58 (.01)	525	670	.75 (.01)

Notes. SDQ: Strengths and Difficulties Questionnaire; Comm.: Community sample; Clin.: Clinical (sub)sample; AUC: Area Under the Curve

[a] Per SDQ scale, the clinical subsamples consisted of adolescent with a DSM-IV diagnosis content-wise matching the SDQ scale: Anxiety/Mood disorder for the SDQ emotional difficulties scale, Conduct / Oppositional Defiant Disorder for the SDQ conduct difficulties scale, Attention-Deficit/Hyperactivity Disorder for the SDQ hyperactivity/inattention difficulties scale and Autism Spectrum Disorder for the SDQ social difficulties and prosocial behaviour scales. For the SDQ total scale, the total clinical sample was used.

DISCUSSION

The aim of this study was to investigate validity aspects of the self-report and parent-report SDQ versions among 12- to 17-year old Dutch adolescents in a community setting. We focused on the SDQ versions' internal structure, and convergent, discriminant, and criterion validity.

Internal structure. Holding ESEM models in higher regard than CFA models, due to the plausibility of items loading on more than one factor, we found some support for the

presumed five-factor structure. However, three items of the SDQ self-report version and six items of the parent-report version were found to be somewhat questionable indicators of their theoretical construct, with one (parent-report version) or two (self-report version) items failing to substantially contribute to the scale they were presumed to contribute to and some items unexpectedly contributing to other scales than their presumed scale. Additionally, the analyses revealed the presence of two to four correlated residuals for both SDQ versions that were not intended to exist. Scale score reliabilities were sufficient for the emotional difficulties and hyperactivity/inattention scales of both SDQ versions and for the parent-reported prosocial behaviour scales, but not for the other scales of both SDQ versions. These findings are cause for concern, but can possibly partially be attributed to the fact that the SDQ aims to measure five dimensions of psychosocial functioning with only five items per dimension. The SDQ's briefness, widely considered to be one of its perks, may come at a cost. Additionally, it is worth noticing that the samples used in this study are presumably large enough to obtain accurate results with CFA's. In contrast, ESEM models are substantially less parsimonious and thus require larger samples (Garrido et al., 2018), which warrants some caution with regard to the results of our ESEM analyses.

For the self-report version, our factor structure and reliability findings are in line with findings by Garrido and colleagues (2018), who performed the only other study using ESEM for assessing the SDQ's scale structure. As none of the other investigations into the factor structure of the self-report and parent-report versions are based on ESEM, it is difficult to compare the findings of the current study to other studies. Our reliability findings appear to deviate from previous research, with most previous studies finding higher reliability estimates than we did. However, note that previous studies have used either Cronbach's alphas or ordinal alphas to estimate reliability, which are both suboptimal measures of the reliability of SDQ scores as Cronbach's alpha does not take the SDQ items' ordinal nature into account and ordinal alpha estimates the reliability of the latent continuous variables underlying the observed scores.

Convergent and discriminant validity. Using the CBCL and YSR as gold standards, we found evidence for the SDQ self-report and parent-report versions' convergent and discriminant validity as, in the great majority of cases, each SDQ scale was more strongly associated with its conceptually similar CBCL/YSR scale(s) than with conceptually different CBCL/YSR scales. These findings are in line with our expectations and with findings from previous studies (Van Widenfelt et al., 2003; Vogels et al., 2011). Note that the comparison with findings from previous studies is slightly hampered by the fact that these studies differed to some extent with regard to the CBCL/YSR scales they identified as conceptually similar to the SDQ scales. Besides, two out of the three studies did not compare SDQ scales to conceptually different CBCL/YSR scales, therewith impeding a comparison of our outcomes regarding discriminant validity with previous studies.

Compared to the above mentioned previous studies, our study adds two unique perspectives to the investigation of the SDQ's convergent and discriminant validity. First,

while previous studies only compared the SDQ scales to the CBCL/YSR empirically based syndrome scales, our study additionally compares the SDQ scales to the CBCL/YSR DSM-oriented scales. The DSM-oriented scales result from a top-down approach of grouping items based on their coverage of DSM symptom categories, whereas the empirically based syndrome scales result from a bottom-up approach of applying statistical analyses to group items. As item grouping based on criteria formulated for diagnostic purposes is clinically relevant, we regard the findings regarding the comparison of the SDQ scales with the DSM-oriented CBCL/YSR scales as additional evidence for the SDQ scales' convergent and discriminant validity.

The second perspective that makes our study stand out from previous studies, is that we investigated the SDQ's discriminant validity by comparing SDQ scales to scales of an instrument from a different domain: the IDS-2 from the domain of intelligence tests. We deem this a useful comparison as lack of a shared domain can be expected to result in weak to negligible associations between scales of instruments from different domains. In the current study, this endeavour resulted in additional evidence for the SDQ's discriminant validity as scores on SDQ and IDS-2 scales appeared to be unrelated or weakly negatively related to each other.

To summarize, our findings suggest that the SDQ measures the intended four types of difficulties and does not unintendedly measure other aspects of behaviour or intelligence.

Criterion validity. For *both* SDQ versions, our findings indicate that the SDQ total difficulties scale can be used to distinguish between community and clinical populations, as is in line with conclusions drawn in previous studies (Goodman et al., 1998; Vogels et al., 2011) In other words, in a screening context the SDQ total difficulties scale can be used to indicate whether an adolescent likely belongs to the clinical population or not. Note that when taking into account the adolescents' gender, the adolescent-reported total difficulties scale was found to distinguish sufficiently well for female adolescents but not for males, indicating that the adolescent-reported total difficulties scale can be used to screen for psychosocial problems among female adolescents and that the same scale of the parent-reported version is useful for both males and females.

Regarding the specific SDQ difficulties and strength scales, *both* SDQ versions' emotional problems, conduct problems and hyperactivity/inattention scales appeared sufficiently capable of distinguishing between the community sample and adolescents diagnosed with an Anxiety/Mood disorder, CD/ODD, and ADHD, respectively. For these scales, no gender differences were found. We have not been able to compare our findings to previous research as, to the best of our knowledge, the criterion validity of the SDQ difficulties scales, other than the aforementioned total difficulties scale, has not been investigated previously. Note that perfect distinction between community and clinical (sub)populations cannot be expected as a) in the community population some undetected psychiatric disorders can be expected to be prevalent and b) adolescents in the clinical population do not only receive DSM-IV diagnoses in one of the four categories that are content-wise corresponding to

the SDQ scales. Moreover, the results may be biased to some extent as it is likely that adolescents with worrisome but minor psychosocial problems are underrepresented in our clinical sample as they may not (yet) be referred to mental health care.

Overall, our findings regarding the criterion validity of the SDQ difficulties scales suggest that they can be used to screen for problems related to Anxiety/Mood disorder, CD/ODD, and ADHD among community adolescent populations. Keep in mind that the SDQ was not developed for diagnostic purposes; after the SDQ is used to provide a preliminary indication of potential problems at hand, thorough assessment by clinicians is needed.

For the SDQ *parent-report* version the social problems scale was found to sufficiently distinguish between the community sample and the clinical sample diagnosed with ASD. In contrast, the parent-reported prosocial behaviour scale and both the adolescent self-reported social problems and prosocial behaviour scales appear insufficiently useful this purpose. In other words, the parent appears to be a better informant for ASD than the adolescent, whereby the parent-reported SDQ social difficulties scale is a useful indicator and the prosocial behaviour scale is not.

Limitations

The preceding discussion of the outcomes of our study implies several strengths. Besides advancing previous research in multiple respects, however, the current study is prone to some potential limitations. First, the community sample data used in this study were gathered in two waves, approximately seven years apart. Moreover, the community sample is not fully representative of the population of Dutch adolescents as adolescents with a mother born in the Netherlands (as opposed to a mother born in another country), adolescents with a mother with a medium educational level (as opposed to low or high), and adolescents living in the east and west of the Netherlands were slightly overrepresented in the community sample. Additionally, the sampling strategies resulted in overrepresentation of 13- and 14-year-olds. By handling these data as being representative of the Dutch adolescent community population, we assume that validity aspects do not change over time and do not depend on characteristics such as ethnicity and age. Though we consider these assumptions to be reasonable, we cannot rule out that the small deviations from the population distribution have resulted in slightly biased results.

The second limitation follows from the fact that our community sample contained missing data at two levels: questionnaire level and item level. All adolescents had data available of at least one SDQ version. For a subset of these adolescents, CBCL/YSR and/or IDS-2 data were available. The missingness of the second SDQ version and the CBCL/YSR questionnaires may not be random, but considering the large numbers of questionnaires that are available to us, we expect the outcomes of this study to be minimally affected. The missingness of IDS-2 questionnaires definitely is not random as only a subsample of the adolescents with at least one SDQ version available was approached to complete

the IDS-2. The adolescents in this subsample showed a relatively low average IQ score and are thus IQ-wise not representative of the population of Dutch adolescents. As we do not know whether the way in which the SDQ measures psychosocial functioning differs across lower and average IQ's, this too may have biased our results to some extent. Regarding the relatively small numbers of missing SDQ, CBCL/YSR and IDS-2 data at item level, we expect the potential bias in our outcomes to be minimal.

Conclusion

The SDQ is widely used to screen for psychosocial problems in community settings. In this study, we found some support for the SDQ's intended scale structure (emotional problems, conduct problems, hyperactivity/inattention, social problems and prosocial behaviour). However, both SDQ versions had some questionable indicators, unintended subfactors, and insufficient scale reliabilities, suggesting that the SDQ's presumed scale structure is not fully tenable among adolescents in a screening setting. In contrast, the results also suggest that the SDQ scales, using CBCL/YSR and IDS-2 scales as criteria, measure the intended types of difficulties and do not appear to unintendedly measure other aspects of behaviour or intelligence. Moreover, the results indicate that both adolescent- and parent-reported SDQ scores can be used to distinguish adolescents likely belonging to the clinical population from other adolescents, and that individual scales from both SDQ versions can be used to identify adolescents with specific types of disorders (parent and adolescent: Anxiety/Mood disorder, CD/ODD, ADHD; only parent: ASD). Evidence regarding the SDQ's scale structure warrants some caution for the use of the scales in their current form. However, the evidence regarding the various validity aspects are mostly supportive for the continued use of the self-report and parent-report SDQ versions as currently used for screening in routine well-child care practice among adolescents.

4

Using the Dutch multi informant Strengths and Difficulties Questionnaire (SDQ) to predict adolescent psychiatric diagnoses

This chapter is based on:
Vugteveen, J., de Bildt, A., Hartman, C. A., & Timmerman, M. E. (2018). Using the Dutch multi-informant Strengths and Difficulties Questionnaire (SDQ) to predict adolescent psychiatric diagnoses. *European Child & Adolescent Psychiatry*, 27(10), 1347-1359. https://doi.org/10.1007/s00787-018-1127-y

This paper was awarded the 'Best student paper prize' at the International Test Commission conference (Montréal, 2018).

ABSTRACT

Knowledge on the validity of the Strengths and Difficulties Questionnaire (SDQ) among adolescents is limited but essential for the interpretation of SDQ scores preceding the diagnostic process. This study assessed the predictive and discriminative value of adolescent-reported and parent-reported SDQ scores for psychiatric disorders, diagnosed by professionals in outpatient community clinics, in a sample of 2,753 Dutch adolescents aged 12 to 17. Per disorder, the predictive accuracy of the SDQ scale that is content-wise related to that particular disorder and the SDQ impact scale was assessed. That is, 24 logistic regression analyses were performed, for each combination of DSM-IV diagnosis (4: Attention-Deficit/Hyperactivity Disorder (ADHD), Conduct/Oppositional Defiant Disorder (CD/ODD), Anxiety/Mood disorder, Autism Spectrum Disorder (ASD)), informant (3: adolescent, parent, both) and SDQ scale(s) (2; related scale only, related scale and impact scale). Additional logistic regression analyses were performed to assess the discriminative strength of the SDQ scales. The results show both fair predictive strength and fair discriminative strength for the adolescent-reported and parent-reported hyperactivity/inattention scales, the parent-reported conduct difficulties scale, and the parent-reported social difficulties and prosocial behaviour scales, indicating that these scales provide useful information about the presence of ADHD, CD/ODD and ASD. The SDQ emotional difficulties scale showed to be insufficiently predictive. The findings suggest that parent-reported SDQ scores can be used to provide clinicians with a preliminary impression of the type of problems for ADHD, CD/ODD and ASD, and adolescent reported scores for ADHD.

INTRODUCTION

Adolescence is a developmental period associated with physical change, psychological development and social adjustments while in the process of acquiring independence. The complexity of these coexisting processes leaves adolescents vulnerable to psychiatric disorders (Abela & Hankin, 2008; Balazs et al., 2013; Olfson, Blanco, Wang, Laje, & Correll, 2014). Estimates of the percentage of adolescents that are referred to outpatient clinics for youth mental or social health care vary between 10 and 15% (Olfson, Druss, & Marcus, 2015; Reijneveld et al., 2014). In outpatient clinics, screening questionnaires are often used as part of the diagnostic process for quickly generating a first impression of the problems at hand. Given the large numbers of adolescents and their parents that fill in such screening questionnaires, a continued research focus should be on how their scores can be helpful in the diagnostic process.

The Strengths and Difficulties Questionnaire (SDQ) is currently one of the most widely used screening instruments (Goodman, 1997; Goodman, 1999). The SDQ can be completed by adolescents themselves (aged 11-16) as well as by parents and/or teachers (for children/adolescents aged 4-16). The questionnaire is relatively short and, as its name suggests, focusses on strengths (prosocial behaviour) as well as deficits (hyperactivity/inattention, conduct problems, emotional problems, peer problems). In addition, the SDQ contains an impact scale which, if an adolescent experiences difficulties, can be used to indicate chronicity, distress, and social impairment for the adolescent as well as burden for others. The usefulness of the SDQ can be judged based on the principles associated with evidence-based assessment (Hunsley & Mash, 2007; Youngstrom, 2013). The core idea of evidence-based assessment is to optimize individual assessment to suit the actual needs of the very individual. According to the principles of evidence-based assessment, an instrument can be a useful addition to a test battery if it is predictive of an important criterion (Youngstrom & Frazier, 2013). The SDQ has repeatedly been evaluated from this perspective, considering several psychiatric disorders as the important criterion (Becker, Hagenberg, Roessner, Woerner, & Rothenberger, 2004; Brøndbo et al., 2011; Goodman, Ford, Simmons, Gatward, & Meltzer, 2000; He et al., 2013; Klasen et al., 2000). Because only a few of these studies have specifically focused on adolescents (Becker et al., 2004; Goodman et al., 2000; He et al., 2013), more research on the accuracy of the SDQ for predicting diagnoses in adolescence is warranted. An important theme herein is that adolescence marks a shift towards using the adolescents themselves as informants, possibly combined with their parents, who are also used as informants during childhood, while increasingly less often using the teachers. At the same time, the parents' role as informants on their children's psychiatric problems slowly decreases and, for most types of problems, eventually ceases to exist.

In the two studies that we could trace in which a comparison was made between adolescent self-report and parent-report (Becker et al., 2004; Goodman et al., 2000; He et

al., 2013), SDQ scores were used to predict psychiatric disorders in any of three categories, namely Attention-Deficit/Hyperactivity Disorder (ADHD), Conduct/Oppositional Defiant Disorder (CD/ODD) (both also referred to as externalizing disorders) and Anxiety/Mood disorder (also referred to as internalizing disorder). Each category of psychiatric disorders was predicted from the SDQ scale that is content-wise related to that particular category of disorders: the hyperactivity/inattention scale for ADHD, the conduct difficulties scale for CD/ODD, and the emotional difficulties scale for Anxiety/Mood disorder. In a large community sample, Goodman and colleagues (Goodman et al., 2000) found that the parent is a better informant than the adolescent, both for externalizing and internalizing disorders. Parent-report yielded fair sensitivity rates for both externalizing disorders (ADHD: .43, CD/ODD: .40) and internalizing disorders (Anxiety/Mood disorder: .39), whereas adolescent self-report yielded low sensitivity rates for internalizing disorders (Anxiety/Mood disorder: .28) and even lower sensitivity rates for externalizing disorders (ADHD: .12, CD/ODD: .15). Becker and colleagues compared adolescent self-report and parent-report among adolescents in a clinical sample and also found the parent to be a better informant than the adolescent for both externalizing and internalizing disorders, but the reliability of these findings is limited because they were found in a rather small sample.

The current study contributes to knowledge about the construct validity of the SDQ by investigating how well diagnoses for specific psychiatric disorders can be predicted from self-reported or parent-reported SDQ scale scores in a large Dutch clinical sample of 2,988 12- to 17-year-old adolescents referred to a mental health outpatient clinic. In line with earlier studies, we aim to predict ADHD, CD/ODD and Anxiety/Mood disorder from their content-wise related scale (i.e., the hyperactivity/inattention scale, the conduct difficulties scale and the emotional difficulties scale, respectively). Additionally, we aim to predict the presence of these disorders from this content-wise related scale combined with the impact scale. Further, we explore how accurately Autism Spectrum Disorder (ASD) diagnoses can be predicted, considering the social difficulties scale and the prosocial behaviour scale as content-wise related scales, as we presume that these scales could have a some predictive value for ASD. We presume so because these scales are intended to provide a comprehensive first screening of social functioning. We acknowledge the existence and value of more specific and thorough ASD instruments, amongst others the Social Responsiveness Scale (SRS; Constantino & Gruber, 2005) and the Children's Social Behavior Questionnaire (CSBQ; Hartman, Luteijn, Serra, & Minderaa, 2006), that contribute to charting the different aspects of ASD. However, such narrow-band instruments only measure ASD; they are different from broad-band screeners such as the SDQ that cover multiple types of psychopathology. In line with previous findings (Becker et al., 2004; Goodman et al., 2000; He et al., 2013), we hypothesize that diagnoses for both externalizing disorders (i.e., ADHD and CD/ODD) and internalizing disorders (Anxiety/Mood) will be predicted fairly accurately. Based on findings from Goodman

and colleagues (Goodman et al., 2000) and general findings from psychopathology research among adolescents (Cantwell, Lewinsohn, Rohde, & Seeley, 1997; Vazire, 2010), we hypothesize the parent to be a better informant than the adolescent for externalizing disorders. Concerning internalizing disorders, we hypothesize the adolescent to be the better informant. We do so because internalizing disorders are considered to be less overt and thus less easily observable by others than by the adolescents who have privileged access to their emotional difficulties such as feeling persistent sadness. This hypothesis is in line with findings from general psychopathology research (Cantwell et al., 1997; Vazire, 2010), but deviates from findings by Goodman and colleagues (Goodman et al., 2000) which suggest that the parent is the best informant for internalizing disorders too. As Goodman's findings were derived from a community sample instead of a clinical sample as is the case in our current study, we base our hypothesis for internalizing disorders on general psychopathology literature. Regarding the prediction accuracy for ASD we expect that parents are better informants than adolescents themselves. This hypothesis is based on the fact that self-report relies on the ability to recognize and verbalize emotions, intentions and functioning, while the limitation in doing so is one of the core symptoms of ASD (American Psychiatric Association, 2013). Additionally, we expect higher levels of adolescent-parent agreement for the externalizing SDQ scales (i.e., hyperactivity/inattention, conduct) than for the internalizing SDQ scale (i.e., emotional difficulties), as is consistent with findings in clinical samples using the Child Behavior Checklist and Youth Self Report (CBCL and YSR, respectively) (Achenbach, & Rescorla, 2001; Rey, Schrader, & Morris-Yates, 1992; Salbach-Andrae, Lenz, & Lehmkuhl, 2009). The SDQ impact scale is not exclusively related to any of the specific types of difficulties that are measured by the SDQ. To our knowledge, the prediction accuracy of the impact scale for specific types of disorders has not been investigated previously and we have no a priori expectations on its predictive strength. In addition to the predictive strength of each scale, we examine its discriminative strength by investigating how well each of the psychiatric disorders are predicted by their non-related scales.

To summarize, the aim of our study is twofold: 1) examine how well specific types of psychiatric disorders, diagnosed in outpatient community clinics, can be predicted from SDQ scales in a large clinical sample of adolescents and 2) investigate whether the accuracy of the prediction depends on the type of informant (adolescent, parent) that was used.

METHODS

Sample

Data were collected from adolescents who had been referred to one of the outpatient clinics of an institution for child and adolescent psychiatry in the North of the Netherlands. The SDQ data were collected online during the intake assessment as part of routine outcome monitoring. The inclusion criteria for the sample were being a first time referral between January 1st of 2013 and December 31st 2015, falling within the age range of 12 through 17 and having received a clinical DSM-IV diagnosis. These criteria were met by 3,826 adolescents. For 2,988 (78.1%) of them, both the self-report and parent-report SDQ data were available. Within this group the mean age was 14.2 years (*SD* = 1.6) among males (54.2%) and 14.6 years (*SD* = 1.5) among females (45.8%).

Missing data. Of the total sample, 838 adolescents were missing SDQ data, from one SDQ informant (adolescent-reported SDQ data missing, *n* = 148; parent-reported SDQ data missing, *n* = 291) or both (*n* = 399). The scores from these adolescents were omitted from the analyses. Table 4.1 provides information about the age, sex and diagnosed disorder distributions within the sample with missing SDQ data (*n* = 838) and within the study sample (*n* = 2,988). The study sample was somewhat younger than the missing data sample ($t(3,826) = 9.20$, $p < .01$, 99% CI [-0.45,-0.69]). Further, in the study sample ADHD diagnoses occurred relatively more frequently, and Anxiety/Mood disorders diagnoses less frequently, than in the missing data sample (ADHD: $z = 4.9$, $p < .01$, 99% CI [0.04,0.13]; Anxiety/Mood: $z = 3.5$, $p < .01$, 99% CI [0.02,0.12]). No evidence was found suggesting that the study sample differed from the missing data sample with respect to gender (male: $z = 1.3$, $p = .20$, 99% CI [-0.03,0.08]) or the prevalence of CD/ODD and ASD (CD/ODD: $z = 2.6$, $p = .01$, 99% CI [-0.01,0.06]; ASD: $z = 1.4$, $p = .15$, 99% CI [-0.02,0.06]).

Table 4.1 Age, sex and diagnosed disorder distributions within the study sample and the missing data sample

	Study sample	Missing data sample
N	**2,988**	**838**
Mean age (*SD*)	14.4 (1.6)	14.9 (1.6)
Male/female	.46/.54[a]	.49/.51
Diagnosed disorders[b,c]		
ADHD	.29	.21
CD/ODD	.11	.14
Anxiety/Mood	.40	.46
ASD	.21	.23

Notes. [a] proportion of N; [b] ADHD: Attention-Deficit/Hyperactivity Disorder, CD/ODD: Conduct/Oppositional Defiant Disorder, ASD: Autism Spectrum Disorder; [c] within both columns, the proportions related to the prevalence of the four disorders add up to more than 1 due to comorbidity of the disorders

Strengths and Difficulties Questionnaire

Dutch translations of the self-report and parent-report SDQ versions were used (Van Widenfelt et al., 2003). The questionnaires consist of 33 items each. The first 25 items cover five scales, with four focusing on difficulties relating to behaviour, emotional functioning, hyperactivity/inattention, interaction with peers, and one focusing on the strength prosocial behaviour. The remainder of the items forms the impact scale which, if an adolescent has difficulties in one or more of the four difficulties domains, can be used to indicate chronicity, distress, social impairment and burden for others. The scales were computed in the standard manner (Goodman, 1997; Goodman, 1999), resulting in scores ranging from 0 to 10 for each scale.

Clinical DSM-IV Diagnosis

The adolescents' clinical diagnoses were established based on thorough diagnostic procedures by trained professionals in a multidisciplinary team, including at least a child- and adolescent psychiatrist, a child psychologist and a specialized nurse. The diagnoses were based on information from various sources. In interviews with the adolescent, current functioning and complaints were assessed, and when assumed relevant, standardized instruments were additionally administered, for example the Anxiety Disorders Interview Schedule (ADIS; Silverman & Albano, 1996), or the Autism Diagnostic Observation Schedule (ADOS; Lord, Rutter, DiLavore, & Risi, 1999; Lord et al., 2012). Parents were interviewed separately from the adolescent about the developmental history of their child, and on current functioning and concerns. Additionally, when assumed relevant, standardized instruments were administered, e.g. the ADIS-P (Silverman & Albano, 1996) or Parent Interview for Child Symptoms (PICS; Schachar & Wachsmuth, 1989). When feasible the teacher(s) of the adolescent was (were) asked to provide information on daily functioning at school and on the adolescent's relationships with adults and peers with the Teacher Telephone Interview for ADHD and related disorders (TTI; Tannock, Hum, Humphries, & Schachar, 2000).

The clinical diagnoses of the sample were grouped into the four DSM-IV categories: ADHD (n = 872, 29.2%), CD/ODD (n = 323, 10.8%), Anxiety/Mood disorder (n = 1,179, 39.5%), Pervasive Developmental Disorder (PDD; n = 620, 20.7%). In this study we use the more current term ASD when referring to PDD. Per DSM-IV category, Table 4.2 provides information about co-occurrence of these diagnoses.

Most notable is the frequent co-occurrence of CD/ODD and ADHD: Of all adolescents with a CD/ODD diagnosis, 46.1% also received a diagnosis from the ADHD category. The other way around occurs much less frequently, as within the ADHD DSM-IV category 17.1% received a CD/ODD diagnosis.

Approximately one out of six adolescents (n = 506, 16.9%) received a diagnosis that did not belong to any of these four categories; 'Eating disorder, not otherwise specified' (n = 182) or 'disorder of infancy, childhood or adolescence, not otherwise specified' (n = 119) were the most frequent.

Table 4.2 Prevalence of comorbidity per DSM-IV diagnosis category

		Comorbid with ...			
DSM category[a]	***N*[b]**	**ADHD[c]**	**CD/ODD[c]**	**Anxiety/mood disorder[c]**	**ASD[c]**
ADHD	872	-	.17	.14	.14
CD/ODD	323	.46	-	.07	.07
Anxiety/Mood disorder	1,179	.10	.02	-	.10
ASD	620	.20	.04	.19	-

Notes. [a]ADHD: Attention-Deficit/Hyperactivity Disorder, CD/ODD: Conduct/Oppositional Defiant Disorder, ASD: Autism Spectrum Disorder; [b]The numbers in this column add up to more than 2,988 (sample size) due to comorbidity; [c]The proportion of adolescents within each DSM category (row), also diagnosed with any of the other disorders

Statistical Analyses

Per disorder, summary statistics (means and standard deviations) were calculated for all SDQ scales for both the self-report version and the parent-report version. Internal consistency information on the SDQ scales for both SDQ versions within in the study sample was retrieved by calculation of Cronbach's alpha coefficients. Per SDQ scale, the Cronbach's alpha coefficients of the two informants were compared with Feldt's test for dependent samples (Feldt, Woodruff, & Salih, 1987).

To assess potential informant effects in combination with disorder effects on SDQ scale scores, a repeated measures multivariate analysis of variance (rm-manova) with the SDQ scale scores as dependent variables and two within-subjects factors (informant and SDQ scale) was conducted for each of the four types of diagnosed disorders.

The strength of the informant agreement between the self-reported and the parent-reported scores per SDQ scale was examined through Pearson's correlations. Differences between correlation coefficients were tested using the Steiger's test (Steiger, 1980).

For comparison with other studies, the ability of the SDQ scales to predict a specific diagnosis was assessed via sensitivity and specificity rates using the 90th percentile score as cutoff score. In the absence of Dutch cutoff scores, we resorted to British population based cutoffs (Goodman, 1997; Goodman et al., 1998).

The predictive value of the SDQ scales for the four disorders considered in this study was assessed by means of logistic regression analysis. These regression analyses were performed for each combination of disorder (4; ADHD, CD/ODD, Anxiety/Mood disorder, ASD) and informant (3; adolescent, parent, both). The predictive value of the SDQ scale content-wise related to the disorder involved was assessed (model 1: SDQ scale as a main effect), as well as the possible additive predictive value of the SDQ impact scale (model 2: SDQ scale and SDQ impact scale as main effects and interaction). This resulted in 24 analyses, all with the probability of receiving a particular disorder versus the probability of receiving any of the other disorders as the outcome.

To account for potential nonlinear relationships between predictor(s) and outcome, we considered the fit of two competing models for each predictor: first a model containing the predictor as a linear effect and second a model containing the predictor as a nonlinear effect via a restricted cubic spline with three knots (Harrell, 2015). From these competing models, the model with the lowest value of Akaike's information criterion (AIC) was retained. The accuracy of the resulting prediction models was assessed with the area under the curve (AUC), corrected for optimism (Efron, 1986), thus expressing the so called outsample prediction value. Using Harrell's guidelines, the optimism of the AUC values was estimated using 500 bootstrap samples (Harrell., 2015). Generally, when AUC values are used to assess predictive strength, values <.70 are considered 'poor', 70-.80 'fair' and ≥.80 'good' (Raiker et al., 2017; Swets, 1988).

We tested the informant effect and model improvement resulting from adding of the impact scale to the models including only the content-wise related SDQ scales with DeLong's method (DeLong et al., 1988). This method can be used to compare AUC values retrieved from correlated models (models 1 or models 2 for predicting a particular disorder based on different informants) and from nested models (models 1 and 2 for predicting a particular disorder based on the same informant).

The discriminative strength of each SDQ scale was investigated by assessing how it predicts the disorders it is content-wise unrelated to. The discriminative strength of a scale is considered fair when the AUC values indicating the prediction accuracy of the scale for all unrelated disorders is <.70, and poor when one or more AUC values ^{3}.70.

For all statistical tests, a significance level of a = .01 was used. All analyses were performed in the R version 3.2.3. (R Core Team, 2016). The logistic regression analyses were performed using the rms package (Harrell., 2017). The comparisons of AUC values were performed using the pROC package (Robin et al., 2011). For illustration purposes, perturbed data and example code are available on https://osf.io/8d7kh/.

RESULTS

Summary Statistics of SDQ Scores

Table 4.3 presents internal consistency information for each of the SDQ scales for the adolescent self-reported and the parent-reported version. Most internal consistency values (Cronbach's alpha) for the SDQ scales range from .71 to .78 and are fairly similar across informants, exceptions being the conduct difficulties scale and the social difficulties scale. For these scales, the internal consistency values with the adolescent as informant are lower (.59 and .55, respectively) than with the parent as informant (.74 and .67, respectively).

Table 4.3 further presents means and standard deviations of SDQ scale scores for both the self-report and parent-report versions, per disorder and across all disorders, with statistics for the content-wise related scale(s) per disorder printed in bold. Column-wise examination of Table 4.3 shows that the highest mean score per scale (and lowest for the prosocial behaviour scale which measures strengths) is found among the adolescents with the corresponding disorder (i.e., the hyperactivity/inattention scale for ADHD; the conduct difficulties scale for CD/ODD; the emotional difficulties scale for Anxiety/ Mood disorder; the social difficulties scale and prosocial behaviour scale for ASD), as was expected. Note that a row-wise examination of the table is not very useful because it only provides a comparison of mean scale scores within a group of adolescents with a particular disorder, thereby ignoring the fact that some types of behaviour are in general less prevalent among patients in outpatient clinics than others. In clinical practice, these differences between types of behaviour are corrected for through the use of cutoff values based on norms (i.e. scores that indicate the level of risk per range of SDQ scale scores) that differ across the SDQ scales.

Comparison of the mean parent and adolescent scores per scale provides an indication of the presence of a potential informant effect on the reported extent of problems. A few exceptions aside, parent-reported mean scores on the SDQ difficulties scales are higher than the equivalent adolescent-reported scores, indicating that parents report a greater degree of difficulties than adolescents. This also holds for the impact these difficulties have on daily life. In the same vein, adolescents report higher prosocial behaviour scores (SDQ strength scale) than parents for all disorders, indicating that adolescents are generally more positive about their strengths than their parents.

Both findings, i.e., 1) the highest (and for the prosocial behaviour scale the lowest) mean score per SDQ scale are found among the adolescents with the corresponding disorder and 2) parent-reported mean scores on the SDQ difficulties scales are generally higher than the equivalent adolescent-reported scores were associated with significant effects on all associated tests in the repeated measures manova.

Informant Agreement

Table 4.4 shows between-informant correlations per SDQ scale across the whole study sample.

The convergent correlations (i.e., correlation between adolescent and parent scores on the same SDQ scale; presented in bold) are positive and range from relatively weak (.34 for impact) to moderately strong (.58 for emotional difficulties). These values indicate limited agreement between adolescents and their parents. A comparison of informant agreement levels on each of the four SDQ difficulties scales revealed no significant differences between the scales, suggesting that adolescent-parent agreement does not depend on the type of problems the informants report on. Compared to informant agreement on the difficulties scales, significantly lower adolescent-parent agreement

was found on the impact scale, suggesting that adolescents and parents more strongly agree on the existence of difficulties than on the impact of difficulties on the adolescents' life. Per SDQ strengths or difficulties scale, the discriminant correlations (i.e., correlations between adolescent and parent scores on different SDQ scales) are significantly and substantially weaker than convergent correlations, which provides evidence for the discriminant validity of the SDQ scales.

In addition to Pearson correlations, we calculated convergent and discriminant intraclass correlation coefficients (available in the Appendix on https://osf.io/8d7kh/). These coefficients show a similar pattern to the one described above.

Table 4.3 Per SDQ version (parent, adolescent) and per SDQ scale: Descriptive statistics and internal consistency information

			SDQ scale						
			Difficulties					**Strengths**	**Impact**
			Hyper	**Conduct**	**Emotional**	**Social**	**Total**	**Prosocial**	**Impact**
		N	**Parent-report**						
Entire sample	Cronbach's alpha[c]	2,988	.76	.74	.75	.67	.78	.75	.71
	M (*SD*)		5.2 (2.8)	2.8 (2.4)	5.2 (2.8)	3.0 (2.3)	16.1 (6.4)	7.4 (2.2)	3.7 (2.6)
ADHD[a]	*M* (*SD*)	872	**7.1 (2.1)**[b]	3.6 (2.5)	4.2 (2.8)	2.6 (2.3)	17.5 (6.3)	7.3 (2.1)	3.6 (2.2)
CD/ODD	*M* (*SD*)	323	6.4 (2.4)	**4.8 (2.4)**	3.8 (2.6)	2.7 (2.2)	17.6 (6.6)	6.7 (2.1)	3.6 (2.4)
Anxiety/Mood	*M* (*SD*)	1,179	4.4 (2.6)	2.2 (2.1)	**6.3 (2.4)**	3.0 (2.2)	15.9 (6.2)	7.7 (2.1)	4.0 (2.7)
ASD	*M* (*SD*)	620	5.6 (2.6)	3.0 (2.5)	5.7 (2.7)	**4.4 (2.3)**	18.6 (6.3)	**6.4 (2.4)**	4.5 (2.5)
			Adolescent-report						
Entire sample	Cronbach's alpha	2,988	.75	.59	.77	.55	.76	.76	.70
	M (*SD*)		5.3 (2.6)	2.5 (1.8)	4.5 (2.8)	2.3 (1.9)	14.7 (5.8)	7.9 (1.8)	2.3 (2.3)
ADHD	*M* (*SD*)	872	**6.8 (2.1)**	3.2 (1.9)	3.4 (2.4)	1.9 (1.7)	15.3 (5.3)	7.7 (1.7)	2.0 (2.0)
CD/ODD	*M* (*SD*)	323	5.7 (2.5)	**3.7 (1.9)**	2.7 (2.1)	2.0 (1.7)	14.1 (5.6)	7.5 (1.8)	1.8 (1.8)
Anxiety/Mood	*M* (*SD*)	1,179	5.1 (2.4)	2.3 (1.7)	**5.9 (2.5)**	2.6 (1.9)	15.9 (5.8)	8.1 (1.7)	3.0 (2.5)
ASD	*M* (*SD*)	620	5.1 (2.6)	2.6 (1.9)	4.3 (2.7)	**3.0 (2.1)**	14.9 (6.1)	**7.4 (2.0)**	2.4 (2.4)

Notes. [a] ADHD: Attention-Deficit/Hyperactivity Disorder, CD/ODD: Conduct/Oppositional Defiant Disorder, ASD: Autism Spectrum Disorder, SDQ: Strengths and Difficulties questionnaire; [b] For each disorder the descriptives of the content-wise related SDQ scale are presented in bold; c Pairwise comparisons of Cronbach alpha coefficients from both informants revealed differences in coefficients for the SDQ conduct ($p < .01$)and SDQ social ($p < .01$) scales, but not for the other scales ($p > .01$)

Table 4.4 Between-informant (adolescent and parent) Pearson correlations per SDQ scale (N = 2,988)

			Parent						
		SDQ scale	**Hyper**	**Conduct**	**Emotional**	**Social**	**Total**	**Prosocial**	**Impact**
Adolescent	Difficulties	Hyper	**.54**[abd]*	.23*	.04	.14*	.35*	.09*	.24*
		Conduct	.37*	**.54***	.01	.15*	.42*	-.26*	.17*
		Emotional	-.18*	-.18*	**.58***	.06	.16*	-.07*	.15*
		Social	-.02	<-.01	.26*	**.54***	.30*	-.08*	.23*
		Total	.26*	.19*	.39*	.31*	**.46***	-.10*	.31*
	Strengths	Prosocial	-.14*	-.19*	.01	-.18*	-.19*	**.41***	-.10*
	Impact	Impact	.03	-.03	.35*	.14*	.20*	.02	**.34**[c]*

Notes. [a] correlations between adolescent and parent scores on the same SDQ scale (convergent correlations) are presented in bold; [b] Pairwise comparisons of the convergent correlations of the four SDQ difficulties scales (bold) revealed no significant differences between the four correlations. For all comparisons: p >.01; [c] Pairwise comparisons of convergent correlations on the SDQ impact scale and each of the SDQ difficulties scales showed that the convergent correlations on the impact scale are significantly lower than convergent correlations on each of the difficulties scales (p <.01); [d] Per SDQ strength or difficulties scale, pairwise comparisons of the convergent correlation (bold) with the discriminant correlations of the remaining strengths and difficulties scales revealed that the convergent correlations were significantly stronger than the discriminant correlations. For all comparisons: p <.01. * correlation significant at the 0.01 level

Predicting Disorders

Table 4.5 presents the sensitivity rate, specificity rate and the diagnostic odds ratio for the content-wise related SDQ scale(s) per type of disorder, using the 90th percentile in British population norms score as cutoff score.

A diagnostic odds ratio larger than 20 characterizes a useful test (Fischer, Bachmann, & Jaeschke, 2003). The diagnostic odds ratios in Table 4.5 range from 2.4 to 5.8, suggesting that the currently used cutoff values may not be appropriate for the clinical population at hand or that the SDQ scales may not be useful predictors. To further investigate the value of the SDQ scales as predictors, a different approach that does not depend on cutoff values might be informative. Such an approach is assessing the SDQ scales' predictive strength through the estimation of prediction models.

Table 4.5 Sensitivity, specificity and the diagnostic odds ratio per SDQ version and disorder based on the British cutoff values (Goodman, 1997; Goodman, Meltzer, & Bailey, 1998)

			Informant					
			Adolescent			Parent		
Disorder category[ab]	**N**	**SDQ scale**	**Sensitivity**	**Specificity**	**OR_D**	**Sensitivity**	**Specificity**	**OR_D**
ADHD	872	Hyper	.59	.76	4.64[c]	.63	.77	5.79
CD/ODD	323	Conduct	.32	.87	3.24	.67	.72	5.14
Anxiety/Mood	1,179	Emotional	.46	.84	4.64	.76	.52	3.52
ASD	620	Social	.12	.95	2.41	.63	.69	3.82
		Prosocial	.09	.97	2.97	.23	.91	3.01

Notes. [a] ADHD: Attention-Deficit/Hyperactivity Disorder, CD/ODD: Conduct/Oppositional Defiant Disorder, ASD: Autism Spectrum Disorder, SDQ: Strengths and Difficulties questionnaire; [b] For each disorder the descriptives of the content-wise related SDQ scale is presented; [c] The diagnostic odds ratio OR_D = (sensitivity x specificity)/((1-sensitivity) x (1-specificity))

Table 4.6 presents the estimated prediction accuracies of two prediction models per disorder, expressed in AUC values. These values indicate how accurately the disorders can be predicted by either the content-wise related scale (model 1) or the content-wise related SDQ scale in combination with the SDQ impact scale (model 2).

The AUC values for the models containing only the content-wise related SDQ scale per disorder (model 1) range from .63 (ASD, adolescent as single informant) to .80 (ADHD, both informants simultaneously), indicating poorly to fairly accurate predictions of the probability of receiving a certain diagnosis. Table 4.6 shows the highest AUC values for ADHD and the lowest for CD/ODD and ASD when the adolescent is used as a single informant and for Anxiety/Mood disorder when the parent is the single informant.

Extending the models with the main effect of the impact scale and its interaction with the content-wise related scale (model 2) improves the accuracy of the prediction for ADHD and CD/ODD (average AUC improvement of .02 and .04 across informants, respectively). For both informants separately and for both informants combined, the change in AUC

values is statistically significant at an α = .01 level. For Anxiety/Mood disorder and ASD, prediction accuracy does not improve when the impact scale is added to the models.

Table 4.6 AUC values (corrected for optimism[a]) for models 1 and 2 per disorder

		Informant[b]			Comparing informants		
		A	**P**	**B**	**AP**	**AB**	**PB**
ADHD[c] (*n* = 872)							
Model 1	hyper	.74[c]	.78	.80	*	*	*
Model 2	incl. impact	.77	.80	.82	*	*	*
CD/ODD (*n* = 323)							
Model 1	conduct	.69	.76	.77	*	*	*ns*
Model 2	incl. impact	.76	.78	.81	*ns*	*	*
Anxiety/Mood disorder (*n* = 1,179)							
Model 1	emotional	.73	.69	.74	*	*	*
Model 2	incl. impact	.73	.70	.75	*	*	*
ASD (*n* = 620)							
Model 1	social + prosocial	.63	.74	.74	*	*	*ns*
Model 2	incl. impact	.64	.74	.74	*	*	*ns*

Notes. [a] Due to the large sample size used in the analyses, the presented optimism-corrected values are equal to the raw AUC values, with the exception of 1) ASD model 1 with the adolescent as informant (raw AUC = .64) and 2) ASD model 2 with both informants (raw AUC = .75); [b] A: adolescent, P: parent, B: both adolescents and parents; [c] ADHD: Attention-Deficit/Hyperactivity Disorder, CD/ODD: Conduct/Oppositional Defiant Disorder, ASD: Autism Spectrum Disorder; * difference between informants significant at the 0.01 level, ns: not significant.

Informant Effects per Disorder

To assess potential informant effects per disorder, a comparison per model (i.e., models 1 and 2) was made between the predictive values (see AUC values and statistical tests in Table 4.6) of the models based on only adolescent information, the models based on only parent information and the models based on both adolescent and parent information.

Attention-Deficit/Hyperactivity Disorder. The parent is the best single informant when either model 1 or model 2 is used for the prediction of ADHD. Compared to using either single informant, the prediction accuracy of the models slightly improves when both informants are used simultaneously.

Conduct/Oppositional Defiant Disorder. The parent is the best single informant when model 1 is used to predict CD/ODD, and using both informants does not improve the prediction accuracy. The AUC values for model 2 do not identify either one of the informants to be superior over the other. Using the informants simultaneously leads to a slight increase in prediction accuracy of model 1 when compared to using the adolescent

as informant, but not compared to the parent as single informant. For model 2, the combination of both informants is superior to using either single informant.

Anxiety/Mood disorder. The adolescent is the best single informant, both when model 1 and when model 2 is used to predict Anxiety/Mood disorder. Using both informants simultaneously hardly improves the prediction accuracy of models 1 and 2, but the improvement is significant.

Autism Spectrum Disorder. The parent is the best single informant for the prediction of ASD for both models. Adding the information provided by the adolescent does not seem to improve the accuracy of the predictions based on parent information.

Discriminative strength

Table 4.7 presents how well each disorders is predicted by each of the SDQ scales. The discriminative strength of each SDQ scale can be assessed by examining how well each disorder is predicted by their content-wise unrelated scales.

The SDQ hyperactivity/inattention scale, conduct difficulties scale, social difficulties scale and prosocial behaviour scale each poorly predict the disorders they are not intended to predict well, regardless of the informant that was used. These findings indicate fair discriminative strength for each of these four scales. The SDQ emotional difficulties scale poorly predicts the disorders it was not intended to predict with the parent as informant, and unintendedly fairly accurately predicts CD/ODD with the adolescent as informant. This indicates fair discriminative strength with the parent and poor discriminative strength with the adolescent as informant.

DISCUSSION

The aim of our study was to examine how well specific types of psychiatric disorders, diagnosed in outpatient community clinics, could be predicted from Dutch SDQ scales in a large clinical sample of 12- to 17-year-olds and to investigate whether the accuracy of the prediction depended on the type of informant that was used. Cutoff values are not available for Dutch adolescents. Using the 90th percentile from the British norms (Goodman, 1997; Goodman et al., 1998) as cutoff scores we found sensitivity rates, specificity rates and diagnostic odds ratios that suggested that either the used cutoff values were not appropriate for the clinical population at hand or that the SDQ scales were not useful as predictors for the disorders (ADHD, CD/ODD, Anxiety/Mood disorder, ASD). In the absence of any further indication of appropriate cutoff scores for the Dutch population and knowing that working with cutoff values entails using limited information from SDQ scale scores (as they are divided into a 3-4 categories only), we proceeded to

Table 4.7 AUC values (corrected for optimism) for each SDQ scale per disorder

Disorder	SDQ scale	Informant		
		A[b]	P	B
ADHD[a] (*n* = 872)				
	Hyper	**.74[c]**	**.78**	**.80**
	Conduct	.64	.64	.65
	Emotional	.67	.64	.68
	Social	.58	.57	.59
	Prosocial	.55	.55	.56
CD/ODD (*n* = 323)				
	Hyper	.54	.64	.64
	Conduct	**.69**	**.76**	**.77**
	Emotional	.72	.66	.72
	Social	.55	.54	.55
	Prosocial	.57	.61	.61
Anxiety/Mood disorder (*n* = 1,179)				
	Hyper	.55	.63	.65
	Conduct	.56	.62.	.62
	Emotional	**.73**	**.69**	**.74**
	Social	.57	.53	.58
	Prosocial	.55	.57	.57
ASD (*n* = 620)				
	Hyper	.53	.54	.59
	Conduct	.49	.54	.54
	Emotional	.53	.56	.60
	Social + Prosocial	**.63**	**.74**	**.74**

Notes. [a] ADHD: Attention-Deficit/Hyperactivity Disorder, CD/ODD: Conduct/Oppositional Defiant Disorder, ASD: Autism Spectrum Disorder, SDQ: Strengths and Difficulties questionnaire; [b] A: adolescent, P: parent, B: both adolescents and parents; [c] For each disorder the descriptives of the content-wise related SDQ scale are presented in bold

investigate the predictive and discriminative strength of the SDQ scales by estimating prediction models. For each SDQ scale (hyperactivity/inattention, conduct, emotional, social and prosocial) and per informant (adolescent, parent or both), prediction models were used to investigate the scale's predictive and discriminative strength. A scale's predictive strength was examined by assessing how well the scale predicted the disorder it was content-wise related to. The discriminative strength of each scale was investigated by assessment of how well the scale predicted the disorders it was content-wise unrelated to. As was hypothesized, we found that diagnoses for externalizing disorders (i.e., ADHD and CD/ODD) and internalizing disorders (Anxiety/Mood) could be predicted fairly accurately from their content-wise related SDQ scale(s), which are the SDQ hyperactivity/inattention scale, conduct difficulties scale and emotional difficulties

scale, respectively. We further found the parent to be the best informant for externalizing disorders, whereas the adolescent was the best informant for internalizing disorders, as is consistent with our hypothesis that was based on general findings from psychopathology research among adolescents (Cantwell et al., 1997; Vazire, 2010). That is, our findings indicate fair predictive strength for the SDQ hyperactivity/inattention scale regardless of the informant that was used. Further, the findings show fair predictive strength for the conduct difficulties scale with the parent as informant and the emotional difficulties scale with the adolescent as informant. Similar levels of adolescent-parent agreement were found across the difficulties scales, which is in contrast with our hypothesis on higher levels of agreement for the externalizing SDQ scales (i.e., hyperactivity/inattention, conduct difficulties) compared to the internalizing SDQ scale (i.e., emotional difficulties). A possible explanation for this deviation is that the group of adolescents with a diagnosis for Anxiety/Mood disorder in our sample consists of relatively many adolescents with anxiety problems (59.5%), few with mood problems (26.4%) and some with both (14.2%). Previous research suggests that, although both regarded as internalizing disorders, anxiety is more easily observable than mood problems (Martel, Markon, & Smith, 2017). Anxiety might therefore not only be relatively accurately reported by the adolescent but also by the parent, resulting in a higher level of adolescent-parent agreement. Regarding the possible additional value of including the impact scale, we did not state a hypothesis. We found that prediction accuracy for only ADHD and CD/ODD disorders improved when the impact of problems was included in the prediction models. This suggests that the impact scale contributes to the prediction of externalizing but not internalizing disorders within a clinical population of adolescents.

Compared to other studies that assessed the SDQ's predictive abilities among adolescents, our study is the first in its attempt to predict ASD from the SDQ. It remains unclear why Goodman (2000), He (2013), Becker (2004) and their respective colleagues refrained from doing so in their studies among adolescents, but in another study (involving children and adolescents without distinguishing between the two) Goodman offers an explanation for omitting patients with ASD: "Firstly, the SDQ is clearly focused on common forms of psychopathology and does not include the sorts of questions that would allow the recognition of autistic or psychotic disorders with confidence. Secondly, it is generally easy to recognize children at risk of psychosis or autism from the referral letter, so there would be little additional merit in predicting these disorders from prior SDQs even if this were possible. Thirdly, new referrals with these disorders are relatively rare in district clinics…" (Goodman, Renfrew, & Mullick, 2000). We only partially agree with Goodman. ASD is a relatively common disorder, with an estimated prevalence up to 1.5% in the general community (Christensen et al., 2016). In our study, no less than 20.7% of the total sample had received an ASD diagnosis. However, characteristics of adolescents referred to outpatient clinics may differ from adolescents in district clinics, which was the setting of Goodman's study. That the adolescents in the current sample

possibly managed to function well enough to avoid an earlier referral, suggests that the adolescents with ASD in our sample were relatively high-functioning. Our sample is therefore not fully representative of the population of adolescents with ASD. Although there is no SDQ scale that is specifically designed to measure autistic behaviour, which was mentioned by Goodman and colleagues as one of the reasons not to include ASD in their study, ASD is defined by social problems. Our findings suggest that, with the parent as informant, ASD can be fairly accurately predicted from the SDQ social difficulties and prosocial behaviour scales, indicating fair predictive strength of these SDQ scales combined. Thus we conclude that for high functioning adolescents with ASD, parent-reported social difficulties and prosocial behaviour can serve as a fairly accurate first-impression proxy of the potential presence of ASD.

To be useful for assessment purposes, the SDQ scales should not only be predictive of the disorder they are content-wise related to, but they should also be able to discriminate between disorders. All but one of the SDQ scales showed fair discriminative strength, regardless of the informant that was used. The exception was the emotional difficulties scale. Although discriminating fairly well with the parent as informant, the emotional difficulties scale did not with the adolescent as informant. Based on the adolescent-reported emotional difficulties scale, CD/ODD was unintendedly predicted fairly accurate. These findings could indicate that the SDQ emotional difficulties scale with the adolescent as informant is of limited use. However, it might be that Anxiety/Mood disorders are underdiagnosed among adolescents with CD/ODD in the sample used in this study. Literature suggests the rates of comorbid Anxiety/Mood disorders among youth with CD/ODD disorders are approximately 40 percent (Greene et al., 2002), whereas this specific type of comorbidity was only found in 7 percent of adolescents with CD/ODD in the sample under study here. If CD/ODD would be indeed underdiagnosed in the current sample, this could explain why the adolescent-reported emotional difficulties scale predicts CD/ODD. The parent-reported emotional difficulties scale, does not appear to be predictive for CD/ODD, possibly because the parent showed to be a poorer informant for emotional problems.

Considering SDQ scales that show both fair predictive strength and fair discriminative strength as useful scales for providing clinicians with a preliminary impression of the type of problems at hand, we conclude that the SDQ hyperactivity/inattention scale is useful for providing information about the potential presence of ADHD, regardless of the informant that was used. The SDQ conduct difficulties scale and the combination of the SDQ social difficulties and prosocial behaviour scales are useful for indicating the presence of CD/ODD and ASD, respectively, with the parent as informant. With the adolescent as informant these scales' predictive strength is inadequate. Further, the SDQ emotional difficulties scale is not useful for assessment as it is not sufficiently discriminative with the adolescent as informant and not sufficiently predictive with the parent as informant; the combination of the adolescent and the parent as informants, does not provide a solution. Consistent with previous research, we investigated informant agreement through the

calculation of correlations between adolescent and parent scores per SDQ scale. We found similar levels of adolescent-parent agreement for externalizing and internalizing difficulties scales. This finding deviates from our hypothesis, which was based on earlier findings that adolescent-parent correlations were higher for externalizing scales than for internalizing scales (Rey et al., 1992; Salbach-Andrae et al., 2009). For various reasons, many studies do not proceed after investigating informant agreement. We strongly recommend to additionally study the association between both adolescent- and parent-reported scores and the diagnosis which the adolescents received, because without it, informant agreement is limitedly useful as it does not provide information about which, if any, of the informants is a good informant. To that end, we performed logistic regression analyses, which identified a best informant for each disorder, with the exception of Anxiety/Mood disorders.

The findings of the current study emphasize the need for an assessment method that combines scores from SDQ difficulties and strength scales with the SDQ impact scale and, as literature on evidence-based assessment suggests too (Hunsley & Mash, 2007), combines information provided by multiple informants. To be optimally useful in clinical practice, this method should result in a probability prediction per type of disorder for each individual. In our view, the methods that are currently most widely used in clinical practice do not fully suffice. That is, using cutoff values results in a categorization into one of three or four (depending on the cutoff solution used) categories per person per SDQ scale. It does not allow combining information from multiple SDQ scales or informants. The alternative is utilizing the algorithm proposed by Goodman and colleagues (Goodman et al., 2000), which combines SDQ difficulties scales with the impact scale and combines information from informants and results in a blunt 'unlikely', 'possible' or 'probable' rating per person per disorder (emotional, conduct or hyperactivity disorder). This method requires information from all informants (adolescent, parent and teacher), which limits its applicability in clinical practice. A useful alternative, would be to use a nomogram (Kattan & Marasco, 2010) derived from a prediction model, estimated based on both community samples and clinical samples. A nomogram is a visual tool that allows the clinical user to retrieve an individual's probability of receiving a particular diagnosis. This tool also visualizes effect sizes per predictor and how predictors interact with each other in predicting the probability of receiving one of the types of disorders.

Strengths and limitations

This study focuses on the validity of the SDQ within a clinical setting. Compared to previous studies among adolescents in a clinical setting, our clinical sample is large and the sample size per disorder is considerable. In that respect, our study clearly surpasses previous studies. Note that our findings pertain to a clinical population and hence do not allow us to infer that the SDQ is useful for detection of psychosocial problems in the general population.

The clinical diagnoses that were predicted in this study, were established by a

multidisciplinary team of trained professionals, based on thorough diagnostic procedures. During these procedures, information was gathered from the adolescents, their parent(s) and, if deemed necessary, their teacher. We realize that this process was not compliant with the STAndardized Reporting of Diagnostic assessment guidelines (STARD) (Bossuyt et al., 2003; Bossuyt et al., 2015), because the diagnoses were only partially corroborated with standardized diagnostic instruments and can thus not be regarded as standardized diagnoses. Besides, literature shows limited agreement between clinician-generated diagnoses and diagnoses generated from standardized procedures (Jensen-Doss, Youngstrom, Youngstrom, Feeny, & Findling, 2014; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009), indicating that the reliability of diagnoses used in this study is potentially limited. However, the clinician-generated diagnoses in this study can be regarded as 'true' in the sense that these were the actual diagnoses that elicited a certain type of treatment. While the use of instruments in a fully standardized procedure, an approach frequently employed in scientific studies, is presumably more reliable, it does not fully represent clinical practice. Given the fact that the clinical diagnoses that were used in the current study are not beyond any doubt, we feel inclined to advocate some cautiousness interpreting the results of our study.

The SDQ data were collected at the start of the diagnostic process as part of the Routine Outcome Measurement (ROM). The ROM data are primarily collected for insurance and policy making purposes. These data were accessible to the multidisciplinary team during their assessment of the adolescents functioning, which is in conflict with the aforementioned STARD guidelines, but typically the data are not used for diagnostic considerations. This is actually one of the main reasons why we conducted the current study, i.e. given that adolescents and their parents spend time filling in this questionnaire, we wanted to provide a thorough evaluation of whether and, if so, how this information can be put to use for their benefit. Hence, though it cannot be ruled out that the ROM data might have influenced the outcome of the diagnostic process for some adolescents, we expect the actual influence of the SDQ scores on the clinical diagnosis to be negligible.

Both limitations just discussed (i.e., the absence of a fully standardized assessment procedure and the accessibility of SDQ information during assessment) might have had an effect on the predictive value of the SDQ scales. The potential effects are in opposite direction. First, the use of clinically-generated diagnoses may have tempered the effects that were found in this study, because a more reliable outcome measure could potentially have been more accurately predicted. Second, the accessibility of the SDQ information during the health care professional's assessment may have affected some of the diagnoses assigned by the professionals, consequently leading to overestimation of the SDQ scales' predictive abilities. As we have no way to estimate the size of these effects, we do not know their net direction and size.

In our study we took comorbidity of disorders into account by considering all registered diagnoses per adolescent. Further research is needed to investigate if combinations of SDQ scales can be used to predict specific types of comorbidity. Additionally, it could be informative to further consider the heterogeneity within a group with a specific disorder. As far as we could trace, all previous studies that assess the SDQ's predictive validity – including ours – investigated how well disorders can be predicted from one or more SDQ scales. By doing so, we neglect the fact that most adolescents with, for instance, an Anxiety/Mood disorder score relatively high on the SDQ emotional difficulties scale, but not all of them score equally high or low on the other SDQ scales. For clinical practice, it could be highly useful to identify SDQ score profiles and investigate how well these profiles predict types of diagnosis. In other words, the next step would be to take diversity in SDQ scores as a starting point and then predict diagnoses as opposed to examining what adolescents with a specific diagnosis have in common, as has been done so far. Such profile information can, as was suggested before, be used to estimate an individual's probability at each of the four types of diagnoses (Kattan & Marasco, 2010).

Implications

Clinical assessment is aimed at diagnosing and planning treatment. It is important that the outcome of clinical assessment is accurate, because the stakes are high for individuals in need of care. Therefore, it is important that assessment is thorough and that only useful tools are used. Considering the SDQ as such a potentially useful tool, we investigated the extent to which Dutch SDQ scales can be used to predict diagnoses and how well they discriminate between different types of diagnoses. The results of this study show that for adolescents referred to an out-patient clinic the SDQ hyperactivity/inattention scale is useful for providing information about the potential presence of ADHD, regardless of the informant that was used. The parent-reported SDQ conduct difficulties scale and the combination of the parent-reported SDQ social difficulties and prosocial behaviour scales are informative about the presence of CD/ODD and ASD, respectively. The SDQ emotional difficulties scale is insufficiently indicative of the presence of Anxiety/Mood disorders, regardless of the informant that was used. It is important to notice that even the most accurate predictions based on the SDQ scales are far from perfect and cannot replace thorough clinical assessment. Additionally, we caution that it is not informative to compare SDQ scale scores within a single individual in order to gain insight into their relative problem levels because some types of behaviour are generally less prevalent or less common than the others. This holds for the general population as well as for (specific) out-patient populations. That, for example, makes a scale score of six (relatively high) on the conduct difficulties scale incomparable to a scale score of six on the hyperactivity/inattention scale (only moderately high). Community-based cutoff values or normed scores may be used for cross-scale comparisons.

The results of this study suggest that it is useful for clinicians to take the SDQ scales, except the SDQ emotional difficulties scale, into account as a first step in the diagnostic process to possibly steer attention towards one or more specific types of disorders, which should then be more thoroughly considered by clinicians. The parent showed to be a useful informant for ADHD, CD/ODD and ASD, and the adolescent for ADHD. For clinical practice, in which it is often challenging to get both the adolescent and the parent to fill in a questionnaire, these findings suggest that it is most useful to ask the parent to fill in the SDQ.

5

The combined self-reported and parent-reported Strengths and Difficulties Questionnaire (SDQ) score profile predicts care use and psychiatric diagnoses

This chapter is based on:
Vugteveen, J., de Bildt, A., Hartman, C., Reijneveld, S.A ., & Timmerman, M.E. (Accepted). The combined self- and parent-rated SDQ score profile predicts care use and psychiatric diagnoses. *European Child & Adolescent Psychiatry.*

ABSTRACT

The Strengths and Difficulties Questionnaire (SDQ) is widely used, based on evidence of its value for screening. This evidence primarily regards the single informant total difficulties scale and separate difficulties subscales. We assessed to what degree adolescents' SDQ profiles that combined all self-reported and parent-reported subscales were associated with use of care and psychiatric diagnoses, and examined the added value thereof over using only a single informant and the total scale. Cluster analysis was used to identify common SDQ profiles based on self-report and parent-report among adolescents aged 12 to 17 in mental healthcare (n = 4,282), social care (n = 124), and the general population (n = 1,293). We investigated associations of the profiles with 'care use' and 'DSM-IV diagnoses', depending on gender. We identified six common SDQ profiles: five profiles with varying types and severities of reported difficulties, pertaining to 95% of adolescents in care, and one without difficulties, pertaining to 55% of adolescents not in care. The types of reported difficulties in the profiles matched DSM-IV diagnoses for 88% of the diagnosed adolescents. The SDQ profiles were found to be more useful for predicting care use and diagnoses than SDQ scores reported by the adolescent as single informant and the total difficulties scale. The latter would have resulted in missing 26% to 54% of the adolescents with problems, namely those with reported emotional difficulties and borderline problem severity. These findings advocate the use of combined self-reported and parent-reported SDQ score profiles for screening.

INTRODUCTION

Approximately 15 to 25% of adolescents experience psychiatric problems (Fergusson et al., 1993; Ormel et al., 2015). To receive adequate mental healthcare, these problems need to be effectively detected and diagnosed. To that end, it is recommended that clinicians consider information on the adolescent's psychosocial functioning provided by multiple informants (American Psychiatric Association, 2013), for instance the adolescents themselves and their parents. Ratings from multiple informants are considered complementary, with more informants better reflecting differences in perspective (Achenbach, McConaughy, & Howell, 1987; De Los Reyes & Kazdin, 2005; Vazire, 2010). One way to gather multiple-informant information for the purpose of screening for psychosocial problems is to ask the informants to complete a questionnaire, such as the widely used Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997; Goodman, 1999). The SDQ contains five subscales (four related to psychosocial difficulties, and one to strengths) and one total difficulties scale.

The validity of the self-report and parent-report SDQ versions for screening is typically investigated by assessing their usefulness for two purposes. The first is distinguishing between adolescents from general and mental healthcare populations, for which the self-reported (Goodman et al., 1998; Theunissen, de Wolff, & Reijneveld, 2019; Vugteveen, de Bildt, Theunissen, Reijneveld, & Timmerman, 2019) and the parent-reported (Vugteveen et al., 2019) total difficulties scales are considered sufficiently useful. The second purpose is predicting the presence of specific disorders regarded to be content-wise related to the constructs measured by the SDQ (Goodman et al., 2000; Russell, Rodgers, & Ford, 2013) among adolescents from mental healthcare populations. Parent ratings were consistently found to be useful for predicting Attention-Deficit/Hyperactivity Disorder (ADHD), Conduct/Oppositional Defiant Disorder (CD/ODD) (Becker et al., 2004; He et al., 2013; Vugteveen, De Bildt, Hartman, & Timmerman, 2018), and Autism Spectrum Disorder (ASD) (Vugteveen et al., 2018). Findings regarding adolescent ratings varied substantially, some supporting their usefulness for predicting ADHD and CD/ODD (Becker et al., 2004; Vugteveen et al., 2018), but not for ASD (Vugteveen et al., 2018). For Anxiety/Mood disorder, findings on the adolescent and parent ratings were too diverse for meaningful conclusions (Becker et al., 2004; He et al., 2013; Vugteveen et al., 2018). Besides, most studies focused on either the adolescent or the parent as informant, therewith providing limited information to inform clinical practice about the usefulness of the recommended multi-informant ratings for screening. Evidence on the latter is lacking.

An additional peculiarity shared by the available studies described above is that they provide information about single domains of behaviour measured by the SDQ (i.e., about the usefulness of the four difficulties scales separately) or about an adolescent's problem behaviour in general (i.e., total difficulty scale, without distinguishing between the domains) and not on the value of using multi-domain SDQ information for screening.

One weakness of this approach is that considering only the total difficulties scale for distinguishing between the general and mental healthcare populations potentially results in clinicians overlooking groups of adolescents experiencing a single type of problems, as they may not score particularly high on the total difficulties scale. Another weakness of considering the total difficulties scale or the separate difficulties subscales for predicting specific disorders is that it provides limited information about the potential presence of co-occurring disorders. That is, the outcome criterion in studies considering the total difficulties scale was typically the presence of at least one disorder, regardless of their total number and specific type(s). The outcome criterion in studies considering separate difficulty subscales was the presence of one specific type of disorder per subscale. With high comorbidity rates in youth with psychiatric problems (Merikangas et al., 2010), this approach over-simplifies reality, with the consequence that the findings from these studies have needlessly limited relevance for clinicians.

The aim of this study is to surpass the limitations of existing findings by assessing whether using adolescents' SDQ profiles that combine all self-reported and parent-reported SDQ subscales, have added value over using a single informant and the total scale for predicting use of care and psychiatric diagnoses. We will do so by first identifying common SDQ profiles based on self-reports and parent-reports among adolescents aged 12 to 17 in child and adolescent mental healthcare (CAMH), child and adolescent social care (CASC), and the general population (community setting). We selected these populations because they represent populations with relatively many adolescents with one or more psychiatric disorders (CAMH), with various psychosocial problems (CASC), and little to no psychiatric problems (community setting; Nanninga, Jansen, Knorth, & Reijneveld, 2018). Next, we will investigate associations of these SDQ profiles with 'care use' and 'DSM-IV diagnoses' (i.e., ADHD, CD/ODD, Anxiety/Mood, ASD, including co-occurring disorders) among diagnosed adolescents, depending on gender. Exploring the potential presence of a gender effect on the usefulness of SDQ profiles for screening can provide further insight as to how to optimize the use of these profiles in clinical practice.

METHODS

Samples

Data were collected from 5,699 12- to 17-year-old Dutch adolescents and their parents. These adolescents were part of the general population (community setting) or were referred to care (CASC and CAMH settings). Table 5.1 provides demographic information on these adolescents and, for comparison, on the Dutch population (Statistics Netherlands, 2015).

Community setting. The data were collected at schools for secondary education in three waves: 1) in 2009/2010 data were collected from 519 13- to 14-year-old adolescents, 2) between 2011 and 2013 from 331 12- to 17-year-olds, and 3) in 2016/2017 from 443 similarly aged adolescents. For these 1,293 adolescents, adolescent-reported SDQ data (*n* = 452), parent-reported SDQ data (*n* = 69) or both or both (*n* = 772) were available.

CASC setting. The CASC data pertains to 124 12- to 17-year-olds referred to child and adolescent social care, from whom adolescent-reported SDQ data (*n* = 19), parent-reported SDQ data (*n* = 31) or both (*n* = 74) were collected between 2011 and 2013.

CAMH setting. Data were collected from two sources: 1) between 2011 and 2013 from 229 adolescents referred to a mental healthcare provider and 2) between 2013 and 2015 from 4,053 adolescents referred to another mental healthcare provider. For the 4,282 adolescents in this sample, adolescent-reported SDQ data (*n* = 367), parent-reported SDQ data (*n* = 245) or both (*n* = 3,670) were available. In this sample, 2,915 adolescents received a DSM-IV diagnosis (American Psychiatric Association, 2000) in any of the four categories (Anxiety/Mood disorder, CD/ODD, ADHD, and ASD) that content-wise respond to the SDQ subscales (see Table 5.2). The diagnoses were established by trained psychologists/psychiatrists in a multidisciplinary team. Another 635 adolescents were diagnosed with other DSM-IV diagnoses and 732 had no registered diagnosis, because they did not meet the DSM-IV criteria for any disorder.

Table 5.1 Demographic characteristics of the adolescents in the community, CASC and CAMH samples

	Community ***(n = 1,293)***	**CASC** ***(n = 124)***	**CAMH** ***(n = 4,282)***	**Dutch population**
Characteristics	**N (%[a])**	**N (%)**	**N (%)**	**%**
Gender				
Male	623 (48.4)[b]	48 (38.7)	2,006 (47.5)[c]	49.5
Female	664 (51.6)	76 (61.3)	2,218 (52.5)	50.5
Age				
12	99 (7.7)[d]	9 (7.3)	615 (14.4)	16.5
13	354 (27.6)	19 (15.3)	785 (18.3)	16.3
14	336 (26.2)	20 (16.1)	816 (19.1)	16.4
15	191 (14.9)	24 (19.4)	838 (19.6)	16.9
16	178 (13.9)	30 (24.2)	713 (16.7)	16.9
17	126 (9.8)	22 (17.7)	515 (12.0)	17.1
Native country mother				
the Netherlands	1,045 (86.2)[e]	92 (88.5)[f]	201 (94.4)[g]	78.6
Other	168 (13.8)	12 (11.5)	12 (5.6)	21.4
Educational level mother				
Low	258 (24.3)[h]	43 (43.9)[i]	59 (28.1)[j]	23.6
Medium	439 (41.3)	50 (51.0)	109 (51.9)	41.7
High	365 (34.4)	5 (5.1)	42 (20.0)	34.7

Notes. CASC = child and adolescent social care; CAMH = child and adolescent mental health; [a] Percentages computed of valid cases only. [b] Missing: n = 6; [c] Missing: n = 58; [d] Missing: n = 9, exact age unknown, but definitely between 12 and 17 years old; [e] Missing: n = 80; [f] Missing: n = 20; [g] Missing: n = 4069; [h] Missing: n = 231; [i] Missing: n = 26; [j] Missing: n = 4072

The Strengths and Difficulties Questionnaire

The 25 items of the Dutch adolescent- and parent-reported versions of the SDQ are evenly divided over five subscales: one for strengths (prosocial behaviour) and four for difficulties (emotional, conduct, hyperactivity, and social problems) (Goodman, 1997; Goodman, 1999; Van Widenfelt et al., 2003). The total difficulties scale consists of the summed four difficulties subscale scores. The items are rated on a three-point scale (0 = *not true*, 1 = *somewhat true* and 2 = *certainly true*). Five positively worded items belonging to different difficulties subscales are reverse-coded. High scores on the four difficulties subscales and the total difficulties scale, represent a high degree of difficulties; a high score on the prosocial subscale represents a high degree of prosocial behaviour. Table A5.1 (appendices, indicated by A, are available on https://osf.io/nqc3j/) reports mean scale scores and standard deviations per setting (community, CASC, CAMH) and informant (adolescent, parent). The information shows that within the community setting adolescents reported higher severity for most types of difficulties than their parents

did, and weaker prosocial skills. Within the CAMH setting, the opposite was found. The findings regarding both settings are in line with previous research (Becker et al., 2004; Van Widenfelt et al., 2003). Within the CASC setting, adolescents reported lower conduct problem severity than their parents did. No informant differences were found for the remaining subscales.

Table 5.2 Prevalence and comorbidity with other disorders per DSM-IV diagnosis category among 2915 diagnosed adolescents within the CAMH sample

DSM category	Gender	Single diagnosis	Comorbid with ...				Total
			Anxiety/ Mood	CD/ODD	ADHD	ASD	
Anxiety/Mood	All[b]	1,152	-	26	103	111	1,392
	M	*297*	-	*12*	*38*	*51*	*398*
	F	*851*	-	*14*	*63*	*60*	*988*
CD/ODD	All[b]	195	26	-	138	11	370
	M	*128*	*12*	-	*106*	*8*	*254*
	F	*63*	*14*	-	*30*	*3*	*110*
ADHD	All[b]	537	103	138	-	110	888
	M	*361*	*38*	*106*	-	*89*	*594*
	F	*174*	*63*	*30*	-	*21*	*288*
ASD	All[b]	486	111	11	110	-	718
	M	*313*	*51*	*8*	*89*	-	*416*
	F	*167*	*60*	*3*	*21*	-	*251*
Multi-problem[a]	All[b]						46
	M						*35*
	F						*10*

Notes. Anxiety/Mood = Anxiety/Mood disorder; ASD = Autism Spectrum Disorder; CD/ODD = Conduct/ Oppositional Defiant Disorder; ADHD = Attention-Deficit/Hyperactivity Disorder, M = male adolescents; F = female adolescents; [a] Adolescents diagnosed with three or more of the above mentioned disorders; [b] Note that the number of male and female adolescents may not add up to the total number of adolescents because information on gender is missing for 58 adolescents in the CAMH sample.

Statistical analysis

We assessed the degree to which adolescents' SDQ profiles were associated with use of care and psychiatric diagnoses by performing a three-step multilevel mixture analysis (Bolck et al., 2004) in LatentGold (Vermunt & Magidson, 2005) on all available adolescent self-reported and parent-reported SDQ subscales simultaneously, thus assuming the data missing at the informant level as missing at random. The first step in the analysis was to identify clusters of adolescents with common SDQ profiles by estimating multilevel mixture models containing one to eight clusters, all with the five SDQ subscales as ordinal dependent variables, the informant (self, parent) at level 1, and the adolescent

at level 2. The model with the smallest Bayes Information Criterion (BIC) (Schwarz, 1978) value was selected for further analysis. The SDQ profiles found were interpreted using British cutoff scores to classify their adolescent self-reported and parent-reported mean SDQ scale scores as 'normal', 'borderline', or 'abnormal' (Goodman, 1997; Goodman et al., 1998). Informant differences were tested using paired sample t-tests, with $\alpha = .01$ and Bonferroni correction for multiple comparisons per cluster. The second step in the analysis was to retrieve the posterior cluster membership probabilities for the selected model. The third and final step was to relate the SDQ profiles to 1) 'care use', by relating cluster membership to setting (community, CASC, and CAMH) and 2) 'DSM diagnoses' for adolescents receiving CAMH, by relating cluster membership to type of diagnosis (anxiety/mood disorder, CD/ODD, ADHD, ASD, and combinations). For both, the interaction with gender was also included. For illustration purposes, perturbed data and example code are available on https://osf.io/nqc3j/.

The SDQ is considered potentially useful for predicting use of care if a) the SDQ profiles indicating the absence of psychiatric problems are mainly prevalent among adolescents not in care and b) the SDQ profiles indicating presence of psychiatric problems are mainly prevalent among adolescents in care, especially those from the CAMH setting. The SDQ is considered useful for obtaining preliminary indications of the disorders present among adolescents if the reported difficulties in the SDQ profiles match the diagnosed disorders.

For conciseness, only gender differences in profile prevalence estimates ≥20% are reported in Tables 5.3 and 5.4. The remaining gender differences can be found in Tables A5.2 and A5.3 (available on https://osf.io/nqc3j/). Prevalence estimates are not reported for (combinations of) disorders that fewer than 100 adolescents within our CAMH sample were diagnosed with.

RESULTS

Identifying common SDQ profiles

Six clusters (i.e. groups) of adolescents, thus six common SDQ profiles, were identified. Per profile, Figure 5.1 presents adolescent self-reported and parent-reported mean scale scores for the strengths and difficulties subscales and total difficulties scale, and their classification according to the range in which they fell (normal, borderline, abnormal). One group had a profile with all means within the 'normal' range, thus we labelled it the 'no difficulties' profile. Two groups each had a profile with one or two mean subscale scores in the 'borderline' range. We labelled those the 'borderline hyperactivity difficulties' and 'borderline conduct and social difficulties' profiles, based on their affected domains. The remaining three groups each showed a profile containing one or more means in the 'abnormal' range. Based on their affected domains, we labelled them the 'emotional difficulties', 'emotional and social difficulties', and 'overall difficulties' profiles.

To validate the stability of this 6-cluster solution across populations the cluster analysis was performed on the community data and on the CAMH data separately. The resulting profiles (Tables A5.5 and A5.6, available on https://osf.io/nqc3j/) highly resembled the six profiles found in the combined samples.

Identifying adolescents in need of care

Per setting (community, CASC, CAMH), Table 5.3 presents the profile prevalence estimates of the six profiles. Additionally, the CASC and CAMH estimates are combined into estimates for adolescents in care.

Community versus in care. The 'no difficulties' profile was estimated to be 11 times more prevalent among community setting adolescents than among adolescents in care (55% and 5%, respectively). In contrast, the five profiles indicating the presence of at least a single type of difficulties were jointly estimated to be over two times more prevalent among adolescents in care than among community setting adolescents (95% and 45%, respectively). For these five profiles, the main differences between community setting adolescents and adolescents in care were found for the profiles with mean scores in the 'abnormal' range: 'emotional difficulties' (community: 9%, in care: 20%), 'emotional and social difficulties' (community: 4%, in care: 21%), and 'overall difficulties' (community: 1%, in care: 20%).

CASC versus CAMH. Differences in prevalence estimates between CASC and CAMH were found for four of the five profiles indicating the presence of difficulties: The 'borderline hyperactivity difficulties' (CASC: 34%; CAMH 16%) and 'overall difficulties' (CASC: 31%, CAMH: 20%) profiles were estimated to be more prevalent among adolescents receiving CASC, and the 'emotional difficulties' (CASC: 8%; CAMH: 20%) and 'emotional and social difficulties' (CASC: 7%; CAMH 22%) profiles were estimated to be more prevalent among adolescents receiving CAMH.

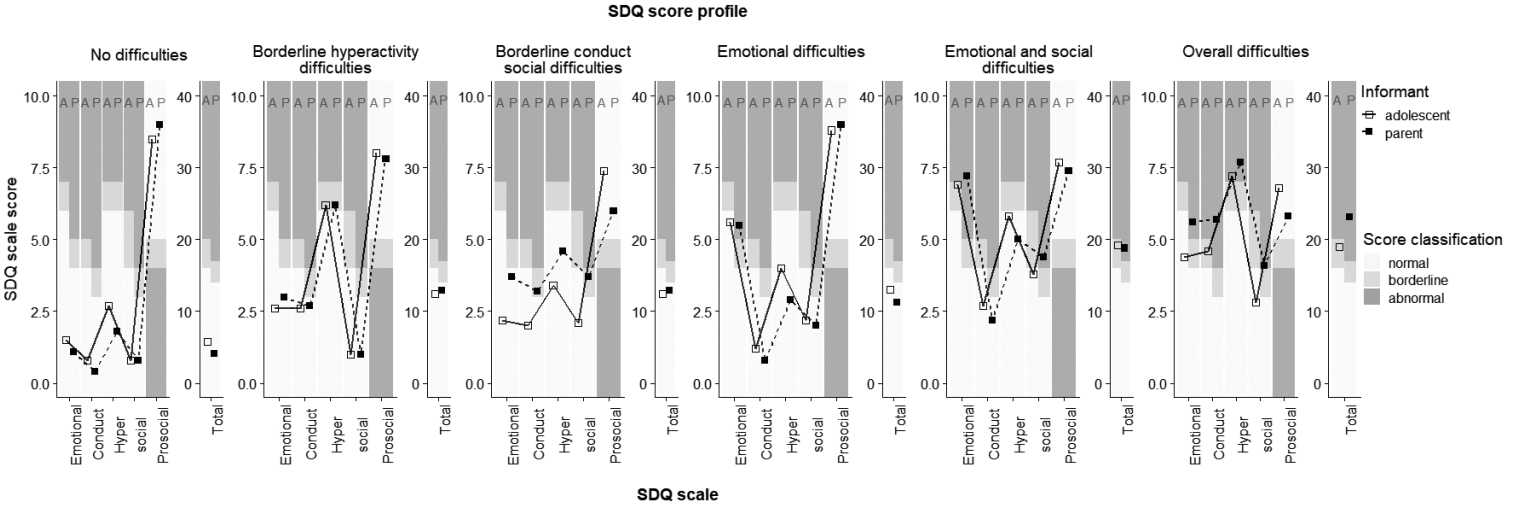

Figure 5.1. Adolescent self-reported (A) and parent-reported (P) mean scale scores per SDQ profile. Table A5.4 (available on https://osf.io/nqc3j/) contains the numerical values of the scale scores presented here.

Table 5.3 Per setting, SDQ profile prevalence estimates in percentages

	SDQ profile					
	No difficulties	**Borderline hyperactivity difficulties**	**Borderline conduct and social difficulties**	**Emotional difficulties**	**Emotional and social difficulties**	**Overall difficulties**
Setting	% All (*M/F*) [a]	% All (*M/F*) [a]	% All (*M/F*) [a]	% All (*M/F*) [a]	% All (*M/F*) [a]	% All (*M/F*) [a]
Community	55	15	17	9	4	1
In care (total)	5	18	16	20 (*8 / 32*)	21 (*11 / 32*)	20
CASC	2	18 (*4 / 27*)	34 (*57 / 20*)	8	7	31
CAMH	5	18	16	20 (*8 / 32*)	22 (*11 / 32*)	20

Notes. SDQ = Strengths and Difficulties Questionnaire; [a] Profile prevalence estimates in percentages for males and females are reported for gender differences >20%

Table 5.4 SDQ profile prevalence estimates per DSM-IV diagnosis (or combination of diagnoses), in percentages of adolescents using child and adolescent mental healthcare (CAMH)

	SDQ profile					
	No difficulties	**Borderline hyperactivity difficulties**	**Borderline conduct and social difficulties**	**Emotional difficulties**	**Emotional and social difficulties**	**Overall difficulties**
DSM-IV diagnosis	% All (*M/F*)	% All (*M/F*)	% All (*M/F*)	% All (*M/F*)	% All (*M/F*)	% All (*M/F*)
Anxiety/Mood	3	5	6	**39**	**38**	**9**
CD/ODD	4	22	**35**	2	3	**33**
ADHD	3	**57 (*65 / 41*)**	2	4	6	29
ASD	2	1	**42 (50 / *26*)**	7	**28**	**21**
Anxiety/Mood & ADHD	1	**20 (*40 / 8*)**	0	**20**	**32**	**26**
Anxiety/Mood & ASD	0	0	**7**	**17**	**66 (*50 / 80*)**	**10**
CD/ODD & ADHD	0	**36**	**6**	0	0	**58**
ADHD & ASD	2	**10**	**21 (*26 / 01*)**	0	**17**	**50 (*44 / 75*)**
Other [b]	10	15	13	36 (*16 / 45*)	16	10

Notes. SDQ = Strengths and Difficulties Questionnaire, M = male adolescents, F = female adolescents. Per disorder (combination), the percentages for content-wise matching SDQ profiles are printed in bold; [a] Profile prevalence estimates for males and females are reported for gender differences >.20; [b] Adolescents diagnosed with DSM-IV disorders other than ADHD, CD/ODD, Anxiety/Mood disorder, ASD

Gender differences. Among adolescents in care, a few gender differences ≥ 20% were found. Males showed a higher estimated prevalence for the 'borderline conduct and social difficulties' (males: 57%; females: 20%) profile within the CASC setting. Females showed higher prevalence estimates for 'borderline hyperactivity difficulties' (males: 4%; females: 27%) within the CASC setting and 'emotional difficulties' (males: 8%; females: 32%) and 'emotional and social difficulties' (males: 10%; females: 32%) within the CAMH setting.

Obtaining a preliminary indication of disorders

For adolescents within the CAMH setting, Table 5.4 presents the prevalence estimates of the six common SDQ profiles per DSM-IV diagnosis, including combinations of diagnoses. Per disorder (combination), the percentages for content-wise matching SDQ profiles are printed in bold. In total, for 88% of the diagnosed adolescents the DSM-IV diagnoses matched the reported types of difficulties.

Anxiety/Mood disorder, and additional diagnoses. As shown in Table 5.4, 86% of adolescents diagnosed with only Anxiety/Mood disorder was estimated to have one of the content-wise matching SDQ profiles ('emotional difficulties': 39%; 'emotional and social difficulties': 38%; 'overall difficulties': 9%). Compared to adolescents diagnosed with only Anxiety/Mood disorder, adolescents with an additional ADHD disorder showed higher prevalence estimates for 'borderline hyperactivity difficulties' (5% versus 20%, respectively) and 'overall difficulties' (9% versus 26%, respectively), and a lower estimate for 'emotional difficulties' (39% versus 20%, respectively). Compared to adolescents diagnosed with only Anxiety/Mood disorder, adolescents additionally diagnosed with ASD showed a higher estimate for 'emotional and social difficulties' (38% versus 66%, respectively) and a lower estimate for 'emotional difficulties' (39% versus 17%, respectively) than adolescents diagnosed with only Anxiety/Mood disorders did.

CD/ODD, and additional diagnoses. Among adolescents diagnosed with only CD/ODD, 68% was estimated to have one of the content-wise matching profiles ('borderline conduct and social difficulties': 35%; 'overall difficulties': 33%). Compared to adolescents diagnosed with only CD/ODD, adolescents additionally diagnosed with ADHD showed higher prevalence estimates for 'overall difficulties' (33% versus 58%, respectively) and 'borderline hyperactivity difficulties' (22% versus 36%, respectively), and a lower estimate for 'borderline conduct and social difficulties' (35% versus 6%, respectively).

ADHD, and additional diagnoses. Among adolescents diagnosed with only ADHD, 86% was estimated to have a content-wise matching SDQ profile ('borderline hyperactivity difficulties': overall: 57 %, males: 65%, females: 41%; 'overall difficulties': 29%). Compared to adolescents diagnosed with only ADHD, adolescents with an additional Anxiety/Mood

diagnosis showed higher prevalence estimates for 'emotional difficulties' (4% versus 20%, respectively) and 'emotional and social difficulties' (6% versus 32%, respectively), and a lower estimate for 'borderline hyperactivity difficulties' (57% versus 20%, respectively). Compared to adolescents diagnosed with only ADHD, adolescents additionally diagnosed with CD/ODD showed a higher estimate for 'overall difficulties' (29% versus 58%, respectively) and a lower estimate for 'borderline hyperactivity difficulties' (57% versus 36%, respectively) than adolescents diagnosed with only ADHD did. Adolescents with an additional ASD diagnosis showed higher estimates for 'borderline conduct and social difficulties' (2% versus 21%, respectively) and 'overall difficulties' (29% versus 50%, respectively), and a lower estimate for 'borderline hyperactivity difficulties' (57% versus 10%, respectively).

ASD, and additional diagnoses. For adolescents diagnosed with only ASD, 91% was estimated to have a content-wise matching SDQ profile ('borderline conduct and social difficulties': overall: 42%, among males: 50%, among females: 26%; 'emotional and social difficulties': 28%; 'overall difficulties': 21%). Compared to adolescents diagnosed with only ASD, adolescents with an additional Anxiety/Mood disorder diagnosis showed a higher prevalence estimate for 'emotional and social difficulties' (28% versus 66%, respectively) and a lower estimate for 'borderline conduct and social difficulties' (42% versus 7%, respectively). Compared to adolescents diagnosed with only ASD, adolescents additionally diagnosed with ADHD showed a higher estimate for 'overall difficulties' (21% versus 50%, respectively) and a lower estimate for 'borderline conduct and social difficulties' (42% versus 21%, respectively) than adolescents diagnosed with only ASD did.

Other or no diagnoses. For adolescents receiving CAMH that are diagnosed with DSM-IV disorders, other than Anxiety/Mood, CD/ODD, ADHD and ASD, the highest profile prevalence estimate was found for the 'emotional difficulties' profile (overall: 36%; among males: 16%; among females: 45%). The probabilities for the remaining profiles were lower and fairly equal to each other (i.e. between 10 and 16%).

Multiple informants versus single informant

Regarding informants, Figure 5.1 and Table A5.4 (available on https://osf.io/nqc3j/) show that the adolescents themselves did not indicate the presence of difficulties for the 'borderline conduct and social difficulties' and the 'emotional difficulties' SDQ profiles, whereas the parents did for one or two difficulties subscales per profile. Based on only adolescent self-report, these two profiles would have merged with the 'no difficulties' profile. This would have resulted in 'no difficulties' being much more prevalent: 81% among adolescents not in care (now 55%) and 41% among adolescents in care (now 5%).

SDQ profiles versus the total difficulties scale

For the groups of adolescents with the 'borderline hyperactivity difficulties', 'borderline conduct and social difficulties', or 'emotional difficulties' profiles, the mean SDQ total difficulties scores were within the 'normal' range. Thus, using the total difficulties scale would have resulted in 'no difficulties' being much more prevalent: 95% among adolescents not in care (now 55%) and 59% among adolescents in care (now 5%).

DISCUSSION

Up to now knowledge was lacking on how the rich information on multiple problem domains captured with the SDQ completed by multiple informants can be used for screening. We addressed this topic by assessing the validity of using adolescents' SDQ profiles that combined all self-reported and parent-reported SDQ subscale information for screening, rather than only separate subscales or total difficulties scores reported by a single informant. Our findings show that the SDQ profile approach is useful for screening, as the profiles were found to be associated with care use, CASC as well as CAMH, and type of diagnosed DSM-IV disorder. Moreover, the SDQ profile approach was found to be more useful for screening than a) a single-informant profile approach, especially if that single informant is the adolescent, and b) using only the total difficulties scale. The validity of using SDQ profiles partly differed for male and female adolescents.

The finding that the SDQ profile approach is more useful for screening than a single-informant profile approach, especially if that single informant is the adolescent, adds in various ways to previous research. Previous research focusing on distinguishing between adolescents from general and mental healthcare populations showed ratings from both informants to be independently useful for this purpose (Goodman et al., 1998; Theunissen et al., 2019; Vugteveen et al., 2019), whereas our findings show that the value of adolescent ratings depends on the type and/or severity of problems present. Moreover, our findings add evidence regarding the unclear value of adolescent self-reported and parent-reported SDQ information for obtaining a preliminary indication of the presence of Anxiety/Mood disorder (Becker et al., 2004; He et al., 2013; Vugteveen et al., 2018) by finding the parent to be an important informant for indicating the presence of Anxiety/Mood disorder. As self-report is commonly regarded as more accurate for internalizing problems (Cantwell et al., 1997; Vazire, 2010), it is a somewhat surprising finding. A potential explanation may lie in the fact that the samples from the CASC and CAMH settings consist of referred adolescents. Our finding could merely reflect the known phenomenon that during adolescence parent-reported need for care exceeds adolescent-reported need for care (Jansen et al., 2013).

The finding that the SDQ profile approach is more useful for screening than only the total difficulties scale, contrasts with previous findings on the value of the total difficulties

scale for distinguishing between adolescents from general and mental healthcare populations. This previous research showed that the adolescent self-reported and parent-reported total difficulties scales were separately useful for that purpose (Goodman et al., 1998; Theunissen et al., 2019; Vugteveen et al., 2019), whereas we found the SDQ total difficulties scale to insufficiently reflect specific psychiatric problems. That is, adolescents whose SDQ subscale scores indicated the presence of emotional difficulties or borderline hyperactivity, conduct and/or social difficulties would have been overlooked based on their total difficulties scale scores. This finding is not surprising, because problems in one or a few domains, especially when it comes to borderline problem severity, do not amount to a substantially increased score on the total problems scale.

In addition, for all types of single DSM-IV diagnoses we found small, yet non-zero prevalence estimates for profiles with types of reported difficulties that did not match the DSM-IV diagnosis involved. We interpret this as an illustration of a well-known phenomenon in informant reports (Gove & Geerken, 1977; Phillips & Clancy, 1970): the intentional or accidental underreporting, overreporting, or misreporting of problems. Although the DSM-IV diagnoses undoubtedly also have errors and partial content overlap and the findings of this study generally support the use of the SDQ for screening, these additional findings emphasize the widely acknowledged limit of using questionnaires as the sole instrument for diagnosing (Smith, 2007).

Implications

Our findings support the combined use of self-reported and parent-reported SDQ subscales for a) distinguishing between adolescents in care and adolescents not in care and b) providing a preliminary indication of the disorders present. We advise against the use of only the SDQ total difficulties scale for screening, as our findings imply that this will result in a substantial number of adolescents with reported problems on the SDQ subscales being overlooked. Our findings further suggest that for screening purposes the parent is more useful as single informant than the adolescent is.

Our exploration regarding gender differences in the validity of using adolescents' SDQ profiles for screening implies that screening accuracy can be improved by applying gender-specific cutoffs for interpreting SDQ scale scores, as internalizing DSM-IV diagnoses were insufficiently reflected in SDQ scores for males, and externalizing diagnoses were insufficiently reflected in SDQ scores for females. It is commonly known that certain behaviours are displayed more frequently or are more outspoken among males than females, and vice versa (Cohen et al., 1993; Merikangas et al., 2010). As this brings about a risk of under-diagnosis of females and males, respectively, we presume that it is of interest to identify adolescents with relatively extreme behaviour compared to other adolescents of the same gender. To facilitate such comparisons, further research is needed to obtain gender-specific cutoff values. The availability of such cutoffs would be consistent with current practice for other questionnaires measuring behaviour, such

as the Child Behavior Checklist (Achenbach, 1991a) and its self-report version the Youth Self Report (Achenbach, 1991b).

Finally, our findings imply that clinicians should be provided with instructions for obtaining probabilities for whether, and if so, which disorder(s) are present. This requires further research, as we could not provide such instructions based on our study, because the samples used are not random samples from their respective populations and we thus cannot estimate the prevalence of the profiles in these populations.

Strengths and limitations

The main strengths of our study are that our findings are based on samples of substantial sizes and that our clinical sample consisted of adolescents with a large variety of mental health problems, yielding a relatively low risk of uncertainty due to sampling fluctuation in our estimations and a relatively high probability that our sample covers the types and severity of problems in the Dutch clinical population. The main limitations of our study are that our samples were not all random samples from their respective populations and were thus potentially not fully representative of the Dutch adolescent populations. Consequently, we do not know how well the profiles we found represent the profiles prevalent in the population. Besides, we used the British cutoff scores to label the profiles, while it is unknown whether they hold for the Dutch adolescent population (Vugteveen et al., 2018). Norms for Dutch adolescents are available (Maurice-Stam et al., 2018; Theunissen, de Wolff, Van Grieken, & Mieloo, 2016), but we refrained from using them as they are based on small samples that are indicated as possibly not representative by the researchers themselves.

Conclusion

This study provides four main insights for the use of the SDQ in practice: 1) the SDQ profiles that combine adolescent self-reported and parent-reported subscale scores are useful for screening, 2) more so than SDQ scale scores reported by a single informant, and 3) more so than using the total difficulties scale. This profile approach can help practitioners put information on multiple problem domains rated by multiple informants to better use for the benefit of adolescents. The usefulness of SDQ profiles for screening can be enhanced by 4) using gender-specific cutoffs, as was indicated by exploratory analyses.

6

Dutch normative data for the self-report and parent-report Strengths and Difficulties Questionnaire (SDQ) for ages 12-17

This chapter is based on:
Vugteveen, J., de Bildt, A., & Timmerman, M. E. (2020*). Dutch normative data for the self-reported and parent-reported Strengths and Difficulties Questionnaire (SDQ) for ages 12-17.* Submitted for publication.

ABSTRACT

The aim of the current study is to present gender-specific and joint normative data per year of age for the Dutch self-report and parent-report SDQ versions for use among 12- to 17-year-old adolescents, based on norm groups consisting of 993 adolescents and 736 parents. We used regression based norming to calculate norms (percentiles and cutoffs) for eight SDQ scales (1 strengths scale, 4 difficulties scales, 1 total difficulties scale, 1 externalizing difficulties scale, 1 internalizing difficulties scale) per SDQ version (adolescent, parent). By design, the gender-specific 'abnormal' cutoffs (i.e., cutoffs identifying maximally 10% of the most extreme scoring males and females, respectively) resulted in about equal percentages of 'abnormal' scoring male and female adolescents per SDQ scale. In contrast, joint 'abnormal' cutoffs (i.e., cutoffs identifying maximally 10% of the most extreme scoring adolescents) resulted in relatively more male (7.6 to 13.6%) than female (3.3 to 8.9%) adolescents as scoring 'abnormal' on scales measuring externalizing behaviour (self-report and parent-report SDQ versions), and relatively more female (3.9 to 14.3%) than male (1.8 to 6.9%) adolescents as scoring 'abnormal' on scales measuring internalizing behaviour (self-report SDQ version). In both types of norms, minor age effects were present. By presenting both gender-specific norms and joint norms, we facilitate the comparison of an adolescent's scores to different reference groups. Besides, the normative data presented in this paper allow for cross-country/ cultural comparisons of adolescents' psychosocial behaviour.

INTRODUCTION

The Strengths and Difficulties Questionnaire (SDQ; Goodman et al., 1998) is widely used to screen for psychosocial problems among adolescents and is valued for several reasons. One is the availability of versions for adolescents themselves, their parent(s) and teacher(s). The availability of these informant versions is essential as it is recommended to gather information from multiple informants for assessing an adolescent's psychosocial behaviour (American Psychiatric Association, 2013). For the SDQ self-report and parent-report versions, ample evidence exists supporting their validity for screening purposes (Becker et al., 2004; Becker et al., 2004; Goodman et al., 1998; Goodman et al., 2000; Goodman, 2001; Lundh et al., 2008; Richter et al., 2011; van Roy et al., 2008; Van Widenfelt et al., 2003; Vugteveen et al., 2018; Vugteveen et al., 2018; Vugteveen et al., 2019). For the teacher version, such evidence is scarce (Becker et al., 2004; Capron, Thérond, & Duyme, 2007).

A second aspect the SDQ is valued for, is its focus on both strengths and difficulties, whereas many other questionnaires only focus on problems. The SDQ consists of one scale measuring strengths (prosocial behaviour) and four scales measuring difficulties (conduct problems, emotional problems, hyperactivity / inattention, peer problems). These four difficulties scales together form the total difficulties scale (Goodman, 1997). Additionally, the conduct and hyperactivity / inattention difficulties scales form the externalizing difficulties scale, and the emotional and social difficulties scales form the internalizing difficulties scale (Goodman, A., Lamping, & Ploubidis, 2010). An individual's SDQ scale scores are typically interpreted using norms, based on the general population. For the SDQ, cutoffs based on these norms are typically determined so that the scores of the ten percent most extreme scoring individuals (scoring high on the difficulties scales, scoring low on the prosocial behaviour scale) are classified as 'abnormal', the scores of the ten percent next-to-most-extreme scoring individuals as 'borderline', and the rest as 'normal' (Goodman, 1997). In other words, the classifications are based on norms corresponding with the 80th and 90th percentiles for the difficulties scales and the 10th and 20th percentiles for the prosocial behaviour scale.

Since the development of the SDQ in 1997, norms were published for the original English SDQ and for several translations. To gain an understanding of how useful these norms are among adolescents, three aspects are important to consider. The first is the availability of age-specific norms. As severity of psychosocial problems is known to be related to age (Costello, Copeland, & Angold, 2011; Durbeej et al., 2019), norms for adolescents should be calculated based on a sample consisting of only adolescents. We found such norms for the parent-reported American (USA) (He et al., 2013), Australian (Mellor, 2005), Chinese (Du, Kou, & Coghill, 2008), Danish (Arnfred et al., 2019), Dutch (Maurice-Stam et al., 2018), Italian (Tobia & Marzocchi, 2018), Israeli (Mansbach-Kleinfeld, Apter, Farbstein, Levine, & Poznizovsky, 2010), Japanese (Moriwaki & Kamio, 2014), Swedish (Björnsdotter, Enebrink,

& Ghaderi, 2013), and Thai (Woerner, Nuanmanee, Becker, Wongpiromsarn, & Mongkol, 2011) SDQ versions, for the self-reported Australian (Mellor, 2005), British (Goodman et al., 1998), Danish (Arnfred et al., 2019), and Israeli (Mansbach-Kleinfeld et al., 2010) versions, and for the teacher-reported Australian (Mellor, 2005), Danish (Arnfred et al., 2019), and Japanese (Moriwaki & Kamio, 2014) versions. Only the norms for the Swedish parent-report version include norms per year of age (10 to 13 years). Across age groups, these norms show differences in percentile ranks corresponding to the SDQ scale scores. This suggests that norms per year of age are more appropriate than norms covering larger age ranges.

The second aspect to consider is the national or geographical background of the individuals in the adolescent sample that the norms were based on. For both the parent-reported (Arnfred et al., 2019; Björnsdotter et al., 2013; Maurice-Stam et al., 2018; Tobia & Marzocchi, 2018) and the self-reported (Arnfred et al., 2019; Goodman et al., 1998) SDQ versions, the SDQ scale score identified as cutoff for the 'abnormal' classification (90th percentile) differed somewhat across language versions, suggesting that norms are potentially of limited use within national, cultural or geographical populations other than the population the norms were determined for.

The third aspect to consider is whether the available norms are gender-specific or not. Gender-specific norms allow for comparing an adolescent's scores to the scores of other adolescents of the same gender. Applying the 'abnormal' cutoffs based on these norms results in identification of the ten percent most extreme scoring adolescents per gender group. In contrast, joint norms allow for comparing an adolescent's scores to those of adolescents in general. Applying the 'abnormal' cutoffs based on these norms results in identification of the ten percent most extreme scoring adolescents, thereby potentially identifying relatively more males than females for some subscales, and vice versa for others. The preference for either gender-specific or joint norms depends on whether SDQ scales measure the intended strengths and difficulties in the same way among male and female adolescents (i.e., whether measurement invariance holds across gender). Joint norms are more appropriate if measurement invariance holds, and gender-specific norms are if it does not. Note that even when a measurement invariance analysis (Millsap & Yun-Tein, 2004) would yield no evidence against measurement invariance, measurement invariance cannot be ruled out. If all items within a scale have a different meaning for boys than for girls, there is no way to distinguish between lack of measurement invariance and difference in means of latent scores across genders. Underlying this gender-specific versus joint norm preference is a debate about a) to what extent the DSM-IV (American Psychiatric Association, 1994) and ICD-10 (World Health Organization, 1992) criteria on which the SDQ items were based, are valid for both genders (Ackermann et al., 2019; Dworzynski, Ronald, Bolton, & Happé, 2012; Mowlem, Agnew-Blais, Taylor, & Asherson, 2019; Waschbusch & King, 2006), b) how stereotypes affect the accuracy of recognizing and reporting an adolescent's problem behaviour by individuals who are key to referral

and diagnostic processes, and c) who needs to be identified with the help of SDQ scale scores (e.g., do we want to identify adolescents who manage to compensate for their symptoms or not?).

For Dutch adolescents, norms based on an adolescent sample are available for the parent-report SDQ version (Maurice-Stam et al., 2018). These norms are neither age-specific nor gender-specific, and they have two additional weaknesses. The first is that the accuracy of these norms may be affected, because the normative sample was potentially not fully representative of the Dutch adolescent population and relatively small (n = 395). Consequently, the resulting cut-off scores may be based on biased norm estimates with substantial uncertainty due to sampling fluctuations. The second weakness of these norms is that they only include norm scores approximately corresponding with the 90th percentile, therewith identifying the 'abnormal' category; norms for identification of borderline cases are lacking. This dichotomization of SDQ scores implies a loss of information, and is arguably less useful for clinical practice. As the Dutch norms for the parent-report SDQ version are potentially of limited use and Dutch norms for other informant versions are lacking, norms for multiple Dutch SDQ informant versions are needed to better facilitate screening.

The aim of the current study is to present gender-specific and joint normative data per year of age for the self-report and parent-report SDQ versions for use among 12- to 17-year-old Dutch adolescents. Norms (percentiles) will be calculated using adolescent samples of decent sizes (self-report: n = 993; parent-report: n = 736), while accounting for potential sample representativity problems regarding gender, socioeconomic status, and ethnic background. Surpassing the methods used in previous SDQ norming studies (i.e., calculating sample percentiles per SDQ scale score per gender/age subgroup), we will estimate population percentiles using regression based (i.e., continuous) norming (Timmerman, Voncken, & Albers, 2019). We will present percentile ranks for all possible scale scores per SDQ scale. Herewith, we facilitate detailed cross-county or -cultural comparisons of SDQ ratings, and provide practitioners with the opportunity to classify an adolescent's score on each SDQ scale without denying them the opportunity to look up the best available estimation of the adolescent's actual percentile score.

METHODS

Norm groups

Data were collected in three waves at schools for secondary education: 1) in 2009/2010 data were collected from 519 13- to 14-year-old adolescents, 2) between 2011 and 2013 from 331 12- to 17-year-olds, and 3) in 2016/2017 from 443 similarly aged adolescents. For 246 of these 1,293 adolescents, information was missing on their age, ethnicity (as indicated by the mother's native country), gender, and/or socioeconomic status (as

indicated by the mother's highest completed level of education). They were excluded from the analyses, as this information was crucial for checking the representativity of the sample. The remaining 1,047 form the norm groups for the self-report (*n* = 993) and the parent-report (*n* = 736) SDQ versions. Table 6.1 provides demographic information on these norm groups and, for comparison, on the Dutch population (Statistics Netherlands, 2015). For the self-report and parent-report SDQ versions, Table 6.2 presents mean scale scores and standard deviations per gender group (males, females) and without distinguishing between genders.

Additionally, Table 6.2 shows per SDQ scale what percentage of adolescents (males, females, total) is identified as scoring in the 'abnormal' range, using previously existing cutoffs. That is, United Kingdom (UK) (Goodman et al., 1998) cutoffs were applied to the self-reported SDQ scale scores; Dutch (Maurice-Stam et al., 2018) and UK (Goodman, 1997) cutoffs were applied to the parent-reported SDQ scale scores. Note that the UK cutoffs for the parent-report version were determined based on a UK sample consisting of both children and adolescents. Because they are not age-specific and from a different country they may be of limited use among Dutch adolescents. With the Dutch norms for the parent-report SDQ version only recently established and such norms for the self-report SDQ version still lacking, these UK norms were widely used in Dutch practice, and still are.

Table 6.1 Demographic characteristic of the adolescents with available SDQ self-reported data (n = 993), with available SDQ parent-reported data (n = 736), and the Dutch population

	SDQ informant version		**Dutch population**
	Self-report	**Parent-report**	
Characteristics	***N (%)***	***N (%[a])***	**%**
Gender			
Male	466 (46,9)	345 (46,9)	49.5
Female	527 (53,1)	391 (53,1)	50.5
Age			
12	82 (8,3)	68 (9,2)	16.5
13	253 (25,5)	186 (25,3)	16.3
14	249 (25,1)	190 (25,8)	16.4
15	151 (15,2)	127 (17,3)	16.9
16	151 (15,2)	97 (13,2)	16.9
17	107 (10,8)	68 (9,2)	17.1
Native country mother			
the Netherlands	885 (89,1)	668 (90,8)	78.6
Other	108 (10,9)	68 (9,2)	21.4
Educational level mother			
Low	238 (24,0)	163 (22,1)	23.6
Medium	411 (41,4)	320 (43,5)	41.7
High	344 (34,6)	253 (34,4)	34.7

Note. SDQ = Strengths and Difficulties Questionnaire

Table 6.2 Per SDQ version (self-report, parent-report), mean scale scores and standard deviations for male and female adolescents

	Gender group								
	Females	% abnormal'[a]		Males	% 'abnormal'		Total	% 'abnormal'	
	Self-report SDQ version								
SDQ scale	M (SD)	UK[b]	NL[c]	M (SD)	UK	NL	M (SD)	UK	NL
Emotional	2.8 (2.3)	96.4	-	1.7 (1.8)	99.6	-	2.3 (2.1)	97.9	-
Conduct	1.2 (1.1)	99.4	-	1.5 (1.4)	98.7	-	1.3 (1.3)	99.1	-
Hyperactivity	3.5 (2.4)	92.8	-	3.8 (2.4)	91.4	-	3.7 (2.4)	92.1	-
Social	1.3 (1.5)	98.5	-	1.4 (1.6)	99.4	-	1.4 (1.6)	99.6	-
Prosocial	8.4 (1.4)	99.1	-	7.7 (1.7)	95.7	-	8.1 (1.6)	97.5	-
Externalizing	4.7 (3.1)	-	-	5.4 (3.2)	-	-	5.0 (3.1)	-	-
Internalizing	4.1 (3.2)	-	-	3.1 (2.8)	-	-	3.6 (3.1)	-	-
Total	8.7 (5.1)	97.2	-	8.5 (4.9)	97.6	-	8.6 (5.0)	97.4	-
	Parent-report SDQ version								
Emotional	2.0 (2.1)	93.6	89.3	1.7 (2.1)	93.0	87.8	1.9 (2.1)	93.3	88.6
Conduct	0.9 (1.2)	98.2	90.0	1.0 (1.5)	96.2	87.0	1.0 (1.4)	97.3	88.6
Hyperactivity	2.1 (2.3)	91.3	93.6	3.3 (2.6)	96.7	86.4	2.7 (2.5)	94.2	90.2
Social	1.3 (1.7)	94.1	94.1	1.7 (1.8)	90.1	90.1	1.5 (1.8)	92.3	92.3
Prosocial	8.5 (1.8)	95.7	86.4	8.1 (1.9)	93.9	81.7	8.3 (1.8)	94.8	84.2
Externalizing	3.0 (3.0)	-	94.4	4.4 (3.5)	-	87.5	3.7 (3.3)	-	91.2
Internalizing	3.3 (3.4)	-	90.5	3.4 (3.3)	-	87.5	3.4 (3.3)	-	89.1
Total	6.3 (5.3)	94.6	92.1	7.8 (5.8)	93.0	87.0	7.0 (5.6)	93.9	89.7

Notes. SDQ = Strengths and Difficulties Questionnaire; [a] The percentage of adolescents with SDQ scores that are considered 'abnormal' using [b] the UK (Goodman, 1997; Goodman et al., 1998) cutoffs (not available for the externalizing and internalizing difficulties scales) and [c] the Dutch (NL) (Maurice-Stam et al., 2018) cutoffs (not available for the SDQ self-reported version).

The Strengths and Difficulties Questionnaire

Adolescents and their parents completed Dutch translations (Van Widenfelt et al., 2003) of the self-report and parent-report SDQ versions, respectively. The 25 items of both versions are evenly divided over five scales: one focusing on strengths (prosocial behaviour) and four scales focusing on difficulties (emotional, conduct, hyperactivity, and social problems). All difficulties items together form the total difficulties scale (Goodman, 1999). Additionally, the conduct problems and hyperactivity / inattention items together form the externalizing difficulties scale, and the emotional and peer problem items together form the internalizing difficulties scale (Goodman et al., 2010). All items are rated on a three-point scale (0 = *not true*, 1 = *somewhat true*, and 2 = *certainly true*). Five positively worded items belonging to different SDQ difficulties scales are reverse-coded. High scores on the difficulties scales represent a high degree of difficulties; a high score on the prosocial behaviour scale represents a high degree of prosocial behaviour.

Statistical analysis

The norm groups were checked for deviations from the Dutch population regarding their distributions of gender, ethnic background (as indicated by the mother's native country) and socioeconomic status (as indicated by the mother's highest completed educational level). Information on the distributions in the Dutch population was retrieved from Statistics Netherlands (Statistics Netherlands, 2015). Note that possible deviations of the norm group distributions regarding age are irrelevant, because we will compute age-specific norms. The information in Table 6.1 indicates that the norm groups for both SDQ versions are not fully representative of the Dutch population of adolescents regarding gender and ethnic background (no problems were detected for socioeconomic status), with an overrepresentation of females and adolescents with a Dutch background. For calculating the joint norms (i.e., without distinguishing between gender groups), the deviations were corrected for by weighing on ethnic background and gender. For calculating the gender-specific norms, the correction was performed by weighing on ethnic background. The weights used are presented in Table 6.3.

Table 6.3 Per SDQ version and type of norms (gender-specific or joint): weights used to correct for oversampling of females and adolescents with a Dutch background

		SDQ informant version			
		Self-report		Parent-report	
Ethnic background	Gender	Gender-specific norms	Joint norms	Gender-specific norms	Joint norms
Dutch	Male	0.45	0.45	0.37	0.37
	Female	0.45	0.40	0.37	0.33
Other than Dutch	Male	1	1	1	1
	Female	1	0.90	1	0.89

Note. SDQ = Strengths and Difficulties Questionnaire

Norms were determined through regression based norming performed in R (R Core Team, 2016), using generalized additive models for location, scale, and shape (GAMLSS package; Rigby & Stasinopoulos, 2005), following the strategy as outlined in Timmerman, Voncken & Albers (2019). Regression based norming allows us to estimate the population distribution of scores per SDQ scale as a continuous function of age (i.e., without splitting up our norm groups into subgroups with certain intervals of age). We opted for this approach because it allows all data to be used simultaneously to establish norms, instead of norms being calculated separately for each subgroup that may or may not be large enough to sensibly perform the necessary calculations on.

Per SDQ version (adolescent, parent), gender-specific norms and joint norms were calculated for 8 scales (1 strengths scale, 4 difficulties scales, 1 total difficulties scale, 1 externalizing difficulties scale, 1 internalizing difficulties scale). Possible scores on the total difficulties scale range from 0 to 40, which can be approximated with a continuous distribution. The population distribution for this scale was estimated using the Box-

Cox power exponential (BCPE) distribution (Rigby & Stasinopoulos, 2004). The BCPE distribution has four parameters: μ for the location of the distribution (median), σ for its scale (approximate coefficient of variation), ν for its skewness (degree of symmetry), and τ for its kurtosis (level of 'peakedness'). The possible scores for the five strengths and difficulties scales (excluding the total difficulties scale) range from 0 to 10, and for the externalizing and internalizing difficulties scales from 0 to 20. These score distributions cannot be reasonably approximated with a continuous distribution. The population distributions for these scales were estimated using the beta binomial (BB) distribution for ordered categorical variables. The BB distribution has two parameters: μ for the location of the distribution (mean) and σ for its scale (approximate coefficient of variation).

In order to calculate the joint norms per year of age (12 through 17), the regression models for all SDQ scales for both SDQ versions included age as predictor for the population distribution parameters (i.e., μ, σ for all scales, and also ν, τ for the total difficulties scale). To consider both linear and more complex associations between age and the distribution parameters, age was included using polynomials. We considered models including polynomials up to degree 20 (i.e., age^1, age^2, …, age^{20}) for each distribution parameter. Per SDQ scale of both SDQ versions (i.e., 16 scales in total), the model with the polynomials resulting in the smallest Bayes Information Criterion (Schwarz, 1978) value was selected. Their fit to empirical data was assessed through visual inspection of worm plots (Buuren & Fredriks, 2001); if needed the models were adapted. The selected models were used to calculate the norms per year of age.

For calculating gender-specific norms, the regression models included both age and gender as predictors for the parameters. Age was included using polynomials, in the same way as for the joint norms. Gender was included as factor as it had two possible values (male, female). Models including the interaction between age and gender were considered. For each of the 16 estimated SDQ scales, the estimated model resulting in the smallest BIC value was selected for visual inspection of its fit, and used (if needed after adaptation) for calculating the gender-specific norms per year of age.

For the sake of conciseness, we present example norms, namely those for 15-year-old male and female adolescents for all scales of the parent-report SDQ version. For all other combinations of age (12 to 17, six ages in total), gender (female, male, total) and SDQ version (adolescent, parent) we present only 'borderline' and 'abnormal' cutoff values. The complete norms can be found in Tables A6.1 through A6.6 (appendices, indicated by A, are available on https://osf.io/4jx8t/). The 'abnormal' cutoffs were established to identify *up to* ten percent of the most extreme scoring adolescents (10th percentile for the prosocial behaviour scale and 90th percentile for all other scales), and the 'borderline' cutoffs were established to identify *up to* ten percent of the next-to-most-extreme scoring adolescents (20th percentile for the prosocial behaviour scale and 80th percentile for all other scales). This approach is in line with how the American (USA) (He et al., 2013), the Australian (Mellor, 2005), and the Danish (Arnfred et al., 2019) norms were determined. In contrast, the Chinese (Du et al., 2008), the pre-existing Dutch (Maurice-Stam et al., 2018), and the

Japanese (Moriwaki & Kamio, 2014) norms were determined to identify *approximately* the percentages mentioned above. For the other adolescent norms, we were unable to determine with certainty which of the two approaches were used. Note that, cutoffs aimed at *approximately* identifying certain percentages can easily be determined based on the information in Tables A6.1 through A6.6 (available on https://osf.io/4jx8t/).

RESULTS

Table 6.4 presents the norms for 15-year-old male and female adolescents for all eight scales of the parent-report SDQ version. Within this age group and for this SDQ version, the norms show higher severity of hyperactivity/inattention and externalizing problems for male than for female adolescents. Consequently, the cutoff values for classifying scores on these scale as 'borderline' or 'abnormal' are higher for males than for females. For example, for females hyperactivity scale scores ≥ 5 are considered 'abnormal', whereas for males scores ≥ 7 are considered as such.

Joint and gender-specific norms

Tables 6.5 and 6.6 present the gender-specific and joint cutoff values per year of age for the self-report and parent-report SDQ versions, respectively. To gain insight into the main differences between the gender-specific norms and the joint norms, we applied the 'abnormal' cutoffs based on both types of norms to the scores of all adolescents in our norm groups. The gender-specific 'abnormal' cutoffs were established to identify a maximum of ten percent of adolescents per gender group, resulting in identification of fairly equal percentages of male and female adolescents as scoring 'abnormal'. In contrast, the joint 'abnormal' cutoffs were established to identify a maximum of ten percent of all adolescents, resulting in the identification of relatively more male than female adolescents as scoring 'abnormal' on scales measuring externalizing problems (self-report and parent-report SDQ versions), and of relatively more female than male adolescents as scoring 'abnormal' on scales measuring internalizing problems (self-report SDQ version). Below these gender differences are described in more detail. The percentages presented can be verified using the cutoffs presented in tables 6.5 and 6.6 in combination with the information in Tables A6.1, A6.2, A6.4, and A6.5 (available on https://osf.io/4jx8t/).

Externalizing problems. For the *self-report* SDQ version, applying the joint 'abnormal' cutoffs resulted in the identification of 10.5% (7.3 to 11.3%, depending on the adolescent's age) of males and 7.7% (5.2 to 8.3%) of females as scoring 'abnormal' on the externalizing difficulties scale. Further, 9.8% (7.6 to 12.1%) of males and 4.4% (3.3 to 5.7%) of females were identified as scoring 'abnormal' on the conduct difficulties scale, and 7.5% (6.6 to 8.4%) of males and 6.6% (5.8 to 7.5%) of females were identified as doing so on the hyperactivity difficulties scale.

Table 6.4 Percentiles for the parent-report SDQ version for 15-year-old males and females

		SDQ scale score																					
SDQ scale	**Gender**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**30**	**40**
Emotional	F	31.9	52.9	68.1	79.0	86.9	92.3	95.9	98.1	99.3	99.8	100												
	M	39.0	58.2	71.0	80.1	86.8	91.7	95.1	97.5	98.9	99.7	100												
Conduct	F	49.8	75.8	89.1	95.4	98.2	99.4	99.8	100	100	100	100												
	M	52.3	73.3	84.9	91.7	95.6	97.9	99.1	99.7	99.9	100	100												
Hyper	F	31.8	52.3	67.1	77.9	85.8	91.5	95.3	97.7	99.1	99.8	100												
	M	14.5	30.7	46.3	60.3	72.2	81.8	89.2	94.4	97.6	99.4	100												
Social	F	43.1	65.2	78.9	87.6	93.1	96.4	98.3	99.3	99.8	100	100												
	M	28.8	52.4	69.8	82.0	90.0	94.9	97.7	99.1	99.7	100	100												
Prosocial	F	100	99.8	99.2	98.0	95.6	91.4	84.5	73.8	57.8	34.4	0.0												
	M	100	99.9	99.5	98.3	95.6	90.2	80.8	66.1	45.6	21.3	0.0												
Externalizing	F	23.2	41.3	55.5	66.7	75.5	82.2	87.3	91.2	94.0	96.0	97.5	98.4	99.1	99.5	99.7	99.9	99.9	100	100	100	100		
	M	12.1	25.7	38.8	50.7	61.1	69.9	77.3	83.2	87.9	91.6	94.3	96.3	97.7	98.6	99.3	99.6	99.8	99.9	100	100	100		
Internalizing	F	20.0	36.6	50.3	61.5	70.6	77.9	83.7	88.2	91.6	94.2	96.1	97.5	98.5	99.1	99.5	99.8	99.9	100	100	100	100		
	M	17.5	33.0	46.3	57.6	66.9	74.7	80.9	85.9	89.8	92.8	95.1	96.8	97.9	98.8	99.3	99.6	99.8	99.9	100	100	100		
Total	F	0.1	11.5	21.7	32.0	42.1	51.4	59.6	66.4	72.1	76.8	80.7	83.9	86.6	88.8	90.6	92.1	93.4	94.4	95.3	96.0	96.7	99.3	100
	M	0.0	7.8	15.1	22.9	31.0	39.1	47.0	54.4	60.9	66.6	71.6	75.8	79.4	82.4	85.1	87.3	89.2	90.8	92.1	93.3	94.3	98.8	100

Notes. SDQ = Strengths and Difficulties Questionnaire; F = females; M = males.

Table 6.5 Cutoff values per year of age for the self-report SDQ version for females, males, and without distinguishing between genders

	SDQ scale															
	Emotional		**Conduct**		**Hyperactivity**		**Social**		**Prosocial**		**Externalizing**		**Internalizing**		**Total**	
Age	**'Bord.'**	**'Abn.'**	**'Bord.'**	**'Abn.'**	**'Bord.'**	**'Abn.'**	**'Bord.'**	**'Abn.'**	**'Bord.'**	**'Abn.'**	**'Bord.'**	**'Abn.'**	**'Bord.'**	**'Abn.'**	**'Bord.'**	**'Abn.'**
	Females															
12	4	6	2	3	6	7	2	3	-[a]	6	7	9	6	8	13	16
13	5	6	2	3	6	7	2	3	6	5	7	9	6	9	13	16
14	5	6	2	3	6	7	2	3	6	5	7	9	7	9	13	16
15	5	6	2	3	6	7	2	3	6	5	7	9	7	9	13	16
16	5	6	2	3	6	7	2	3	6	5	7	9	7	9	13	17
17	5	6	2	3	6	7	2	3	6	5	7	9	7	9	13	17
	Males															
12	3	4	3	4	6	7	3	4	5	4	8	10	5	7	13	16
13	3	4	3	4	6	7	3	4	5	4	8	10	5	7	13	16
14	3	4	3	4	6	7	3	4	5	4	8	10	5	7	13	16
15	3	4	-[a]	3	6	7	3	4	5	4	8	10	6	7	13	16
16	3	4	-[a]	3	6	7	3	4	5	4	8	10	6	7	13	16
17	4	5	2	3	6	7	3	4	5	4	8	10	6	7	13	16
	Total															
12	4	5	-[a]	3	6	7	3	4	6	5	8	10	6	8	13	16
13	4	5	2	3	6	7	3	4	6	5	8	9	6	8	13	16
14	4	5	2	3	6	7	3	4	6	5	8	9	6	8	13	16
15	4	5	2	3	6	7	3	4	6	5	8	9	6	8	13	16
16	4	6	2	3	6	7	3	4	6	5	8	9	6	8	13	16
17	5	6	2	3	6	7	3	4	6	5	8	9	7	8	13	16

Notes. SDQ = Strengths and Difficulties Questionnaire; 'Bord.' = 'borderline'; 'Abn.' = 'abnormal'.
For all SDQ scales except the prosocial scale, scale scores equal to or higher than the displayed cutoff value (≥ 90th percentile) are considered 'abnormal'. For the prosocial scale, scale scores equal to or lower than the displayed cutoff value (≤ 10th percentile) are considered 'abnormal'.
For all SDQ scales except the prosocial scale, scale scores equal to the displayed cutoff value up to the cutoff value for 'abnormal' (80th up to the 90th percentile) are considered 'abnormal'. For the prosocial scale, scale scores scale scores equal to or lower than the displayed cutoff value (20th down to the 10th percentile) are considered 'borderline'.
[a] Borderline cutoff score unavailable due to right skewed score distribution, resulting in large jumps in percentile ranks and skipping the 80th up to the 90th percentiles.

Table 6.6 Cutoff values per year of age for the parent-report SDQ version for females, males, and without distinguishing between genders

	SDQ scale															
	Emotional		Conduct		Hyperactivity		Social		Prosocial		Externalizing		Internalizing		Total	
Age	'Bord.'	'Abn.'	'Bord.'	'Abn.'	'Bord.'	'Abn.'	'Bord.'	'Abn.'	'Bord.'	'Abn.'	'Bord.'	'Abn.'	'Bord.'	'Abn.'	'Bord.'	'Abn.'
	Females															
12	4	5	2	3	5	6	2	4	6	5	6	8	6	8	11	15
13	4	5	2	3	4	6	3	4	6	5	6	8	6	8	11	15
14	4	5	2	3	4	6	3	4	6	5	5	7	6	8	11	14
15	4	5	2	3	4	5	3	4	6	5	5	7	6	8	10	14
16	4	5	2	3	3	5	3	4	6	5	5	7	6	8	10	14
17	4	5	2	3	3	5	3	4	6	5	5	7	6	8	10	13
	Males															
12	4	5	2	3	6	8	3	5	6	5	8	10	6	9	13	18
13	4	5	2	3	6	7	3	5	6	5	8	10	6	9	13	18
14	4	5	2	3	6	7	3	5	6	5	7	9	6	9	13	17
15	3	5	2	3	5	7	3	4	6	5	7	9	6	9	13	17
16	3	5	2	3	5	6	3	4	5	4	7	9	6	8	12	17
17	3	5	2	3	5	6	3	4	5	4	6	8	6	8	12	16
	Total															
12	4	5	2	3	6	7	3	4	6	5	6	10	6	9	13	17
13	4	5	2	3	5	7	3	4	6	5	6	9	6	8	12	17
14	4	5	2	3	5	6	3	4	6	5	6	9	6	8	12	16
15	4	5	2	3	5	6	3	4	6	5	6	8	6	8	11	16
16	3	5	2	3	4	6	3	4	6	5	6	8	6	8	11	15
17	3	5	2	3	4	5	3	4	6	5	5	7	6	8	11	14

Notes. SDQ = Strengths and Difficulties Questionnaire; 'Bord.' = 'borderline'; 'Abn.' = 'abnormal'.
For all SDQ scales except the prosocial scale, scale scores equal to or higher than the displayed cutoff value (≥ 90th percentile) are considered 'abnormal'. For the prosocial scale, scale scores equal to or lower than the displayed cutoff value (≤ 10th percentile) are considered 'abnormal'.
For all SDQ scales except the prosocial scale, scale scores equal to the displayed cutoff value up to the cutoff value for 'abnormal' (80th up to the 90th percentile) are considered 'abnormal'. For the prosocial scale, scale scores scale scores equal to or lower than the displayed cutoff value (20th down to the 10th percentile) are considered 'borderline'.

For the *parent-report* SDQ version, applying the joint 'abnormal' cutoffs resulted in the identification of 11.0% (8.9 to 13.6%) of males and 5.4% (4.2 to 6.9%) of females as scoring 'abnormal' on the externalizing difficulties scale. Further, 8.2% (7.6 to 8.7%) of males and 4.4% (4.1 to 5.0%) of females were identified as scoring 'abnormal' on the conduct difficulties scale, and 11.0% (8.8 to 13.2%) of males and 4.7% (3.7 to 5.9%) of females were identified as doing so on the hyperactivity difficulties scale.

Internalizing problems. For the *self-report* SDQ version, applying the joint 'abnormal' cutoffs resulted in the identification of 5.9% (5.8 to 6.1%) of males compared to 10.7% (10.0 to 11.7%) of females as scoring 'abnormal' on the internalizing difficulties scale. Further, 3.3% (1.8 to 4.2%) of males and 11.6% (8.3 to 14.3%) of females were identified as scoring 'abnormal' on the emotional difficulties scale, and 6.2% (5.6 to 6.9%) of males and 4.5% (3.9 to 5.1%) of females were identified as doing so on the social difficulties scale. For the *parent-report* SDQ version, no substantial gender differences in reported internalizing problems were found.

DISCUSSION

The SDQ is widely used to screen for psychosocial problems among adolescents. Norms for interpreting SDQ scale scores are available for multiple language versions of the questionnaire. However, for none of these language versions joint norms and gender-specific norms per year of age were established, even though the occurrence of psychosocial problems is known to be related to age (Costello et al., 2011; Durbeej et al., 2019) and gender (Merikangas et al., 2010; Vollebergh et al., 2006). We addressed this issue by providing such norms for the Dutch self-report and parent-report SDQ versions for use among 12- to 17-year-old adolescents. The norms showed the presence of age- and gender-effects in reported problem severity.

The Dutch self-report and parent-report SDQ versions were introduced in 2003 (Van Widenfelt et al., 2003), with UK joint norms available for interpreting SDQ scale scores (Goodman, 1999). In 2019, Dutch norms were provided for the parent-report SDQ version (Maurice-Stam et al., 2018). In our norm groups, we found cutoffs based on the UK norms to yield detection rates much lower than the intended ten percent of the most extreme scoring adolescents, especially for the self-report SDQ version. The cutoffs based on the Dutch norms for the parent-report SDQ version yielded varying results, with detection rates close to ten percent for some scales and much lower or higher for other scales. Compared to the pre-existing UK and Dutch norms, we presume our newly established norms to be more useful for interpreting Dutch adolescents' scores because they are a) recent (Evers et al., 2010; Wasserman & Bracken, 2013), b) age-specific, c) available for the self-report and the parent-report SDQ versions, d) established using regression based

(i.e., continuous) norming, and e) based on decent sample sizes, with representativity issues corrected for. Besides, we provide not only joint norms, but also gender-specific norms, therewith facilitating comparison of an adolescent's scores to different reference groups.

Limitations

The validity of the norms presented in this paper is potentially affected by two aspects. The first is our effort to correct for norm group deviations from the Dutch adolescent population regarding ethnic background and gender by applying weights. To the best of our knowledge, this is an acceptable way to deal with these norm group representativity issues that presumably introduced little bias. The second specifically regards the gender-specific norms. In the Dutch language, sex and gender are often indicated with the same word. As this word was used in the questionnaires, the resulting indications can be interpreted as gender and as sex. Calling our norms gender-specific might thus be somewhat inaccurate, as we cannot be sure that gender was provided for adolescents whose biological sex contrasts their gender identity.

Conclusions

This study provides joint and gender-specific norms (percentiles) per year of age for all adolescent self-reported and parent-reported Dutch SDQ scales, including the externalizing and internalizing difficulties scales. The gender-specific norms yield different results than joint norms do. They confirm that females tend to report higher internalizing problem severity and males and their parents tend to report higher externalizing problem severity. By presenting both types of norms, we facilitate the comparison of an adolescent's scores to different reference groups: All similarly aged other adolescents or all similarly aged adolescents of the same gender. Besides, the normative data presented in this paper allow for cross-country/cultural comparisons of adolescents' psychosocial behaviour.

7

General discussion

GENERAL DISCUSSION

The aim of this thesis was to provide knowledge beneficial for optimizing the use of the self-report and the parent-report Strengths and Difficulties Questionnaire (SDQ) versions among Dutch adolescents aged 12 to 17 years in screening and diagnostic procedures in healthcare practice. The findings predominantly supported the use of the SDQ, especially the parent-report version, in these procedures. Dutch gender-specific norms and joint norms for interpreting SDQ scale scores are provided. In the future, the use of the information contained in SDQ scale scores can be optimized by considering an adolescent's SDQ score profile that combines self-reported and parent-reported SDQ scales.

This chapter provides a discussion of the findings that are presented in Chapters 2 to 6 of this thesis, starting with findings related to construct and criterion validity aspects. Next, findings related to scale score reliability will be discussed, followed by a brief discussion of the norms for the Dutch self-report and the parent-report SDQ version. In the remaining part of this chapter, I will reflect on the strengths and limitations of the studies, their practical implications and open issues for future research.

Construct validity

In this thesis, three aspects of construct validity of the self-report and parent-report SDQ versions were assessed. The first is their internal structure in, and measurement invariance across community (e.g., as part of a routine well-child check-up or at school) and clinical (e.g., during intake preceding thorough diagnostic assessment by clinicians) settings. The second is whether known differences between adolescents in these settings are reflected in scores on these SDQ versions. The third is the comparability of SDQ scales to CBCL/YSR scales that are supposed to measure similar constructs, and to CBCL/YSR and IDS-2 scales that are supposed to measure different constructs (i.e., assessing convergent and discriminant validity, respectively).

Internal structure and measurement invariance. Our findings largely supported the intended five-scale structure of the self-report and parent-report SDQ versions, and suggested that measurement invariance across clinical and community settings holds. More specifically, for both SDQ versions the findings reported in Chapters 2 and 3 confirmed that the constructs (prosocial behaviour, emotional difficulties, conduct difficulties, hyperactivity/inattention difficulties, and social difficulties) measured by the five separate strengths and difficulties scales could be distinguished from each other fairly well, that each scale measured mainly one construct, and that all items per scale contributed to measuring that construct.

In line with several previous studies our findings indicated the presence of some deviations from the intended scale structure. That is, the results corroborated that the adolescents' answers were affected by the positive wording of five items belonging

to different difficulties scales (van Roy et al., 2008), suggesting an unintended overlap between the scales of the self-report SDQ version. Additionally, our findings confirmed that a few items per SDQ version were fairly weak indicators of the construct they are supposed to be indicators for (Garrido et al., 2018), and that some scales measured a main construct and one or more subconstructs (van de Looij-Jansen et al., 2011). One example of the latter regards the prosocial behaviour scale. Within this scale, a subset of two items measured a common aspect (i.e., 'helping' behaviour) besides contributing to measuring prosocial behaviour. Though the quality of the SDQ as a screening instrument might be improved when these deviations from the intended internal structure would be solved by adapting the questionnaire, I consider the use of the scales in their current form warranted, as all scales, are left with at least three (parent-reported social difficulties scale) or four (all other scales) proper indicators. Additionally, only one out of the five self-reported positively worded items was found to be a weak indicator for its intended construct. Moreover, the number of subconstructs seemed limited and their presence in some cases seemed to result directly from the fact that SDQ scales measure several related constructs (e.g., the emotional difficulties scale covers anxiety-related difficulties and mood-related difficulties).

Allowing the above-mentioned deviations, the self-report and the parent-report SDQ versions were found to be measurement invariant across community and clinical settings, suggesting that SDQ scores gathered in the two settings bear a similar meaning in terms of problem severity. This means that scores gathered in both settings can be interpreted using the same norms and can be compared to each other.

Known group differences. The findings reported in Chapter 3 suggest that SDQ scale scores reflect the expected difference in problem severity among adolescents from general populations and adolescents from mental healthcare populations (Goodman, 1999). That is, the results in Chapter 3 showed that higher levels of problem severity and weaker prosocial skills were reported for adolescents from a mental healthcare group compared to adolescents from the general population. Adding to findings regarding the SDQ's internal structure and measurement invariance across community and clinical settings, these group difference findings suggest that SDQ scales adequately measure symptom severity on domains of psychosocial behaviour.

Convergent and discriminant validity. The results reported in Chapter 3 suggested that the SDQ measures the intended domains of psychosocial behaviour (i.e., prosocial behaviour, emotional difficulties, conduct difficulties, hyperactivity/inattention difficulties, and social difficulties), as each SDQ scale was more strongly associated with its conceptually similar CBCL/YSR (Achenbach, 1991a; Achenbach, 1991b) scale(s) (i.e., evidence for convergent validity) than with conceptually different CBCL/YSR and IDS-2 (Grob, Hagmann-von Arx, Ruiter, Timmerman, & Visser, 2018) scales (i.e., evidence for discriminant validity).

Criterion validity

In this thesis, criterion validity aspects of the self-report and parent-report SDQ versions were assessed by investigating the value of the SDQ for use in community and clinical settings, with a focus on identifying adolescents at risk of psychiatric disorders that are related to the domains of psychosocial behaviour covered by the SDQ: Anxiety/ Mood disorder, Conduct/Oppositional Defiant Disorder (CD/ODD), Attention-Deficit/ Hyperactivity Disorder (ADHD), and Autism Spectrum Disorder (ASD).

Community setting. The findings reported in Chapter 3 largely supported the use of both SDQ versions, as tools contributing to the identification of adolescents at risk of the above-mentioned psychiatric disorders in a community setting (i.e., mainly healthy adolescents). That is, in line with previous studies (Goodman et al., 1998; Vogels et al., 2011) the total difficulties scales of both SDQ versions were found to be useful for indicating whether an adolescent likely belongs to the clinical population or not, with the adolescent-reported total difficulties scale found to be more useful among females than among males. This gender difference can be attributed to problems in self-assessment of severity of social difficulties among males. The findings in Chapter 3 further suggest that the emotional, conduct, and hyperactivity/inattention scales of both SDQ versions are useful for identifying adolescents (both males and females) with anxiety/mood disorder, CD/ODD, ADHD, respectively, as is the parent-reported social difficulties scale for identifying adolescents diagnosed with ASD.

Clinical setting. The findings reported in Chapter 4 indicated that both adolescent self-reported and parent-reported SDQ scale scores gathered in a clinical setting (i.e., adolescents with on average relatively high problem severity) are fairly useful for providing clinicians with a preliminary indication of the presence of a DSM-IV diagnosis for ADHD. Additionally, parent-reported SDQ scale scores were found to be useful for providing preliminary indications of CD/ODD and ASD. These findings suggest that the parent is a better informant for externalizing disorders than the adolescent themselves, which is consistent with general findings from psychopathology research (Cantwell et al., 1997; Vazire, 2010). Further, SDQ scale scores were found to be insufficiently useful for obtaining an indication of the presence of Anxiety/Mood disorder, regardless of whether adolescent self-ratings or parent-ratings were used.

SDQ score Profile approach in community and clinical settings. The Strengths and Difficulties Questionnaire (SDQ) is widely used, based on evidence that mostly regards separate difficulties scales reported by a single informant. In contrast to using a single scale at a time, the findings reported in Chapter 5 advocate considering SDQ score profiles. These SDQ score profiles combine all self- and parent-reported SDQ scales information for indicating whether an adolescent likely belongs to the clinical population

or not, and for obtaining a preliminary indication of the type(s) of problems at hand. Compared to considering separate scales reported by a single informant, this multi-informant and multi-domain approach does more justice to the complexity of diagnoses, the adolescents' corresponding behaviour, and the high comorbidity rates in youth with psychiatric problems (Merikangas et al., 2010). Consistent with the findings in Chapter 4, the SDQ ratings provided by the parent were found to be more informative than the ratings provided by the adolescents themselves. Additionally, the usefulness of SDQ score profiles was found to depend on the adolescents' gender, suggesting that the use of the SDQ can be improved by applying gender-specific cutoffs. That is, with the current non-gender-specific UK-based cutoffs (Goodman, 1997; Goodman et al., 1998) internalizing DSM-IV diagnoses were insufficiently reflected in SDQ scores for males, and externalizing diagnoses were insufficiently reflected in SDQ scores for females.

To summarize, the adolescent self-reported SDQ version seems more useful among adolescents with on average low problem severity (community setting) than among adolescents with on average high problem severity (clinical setting), whereas the parent-report version was found to be useful among both groups of adolescents.

Reliability

Chapters 2 and 3 present information about the reliability of the SDQ scales in the form of Cronbach's alpha and nonlinear structural equation modelling reliability coefficients. Cronbach's alpha coefficients are lower-bound estimates of scale score reliability, whereas nonlinear structural equation modelling reliability coefficients (Yang & Green, 2015) are estimates of the actual scale score reliability. Consequently, we deem Cronbach's alpha of little interest as a measure of reliability; they were reported only for comparability with other studies. The nonlinear structural equation modelling reliability (ρ_{NL}) coefficients suggested that the conduct and social difficulties scales of both SDQ versions and the prosocial behaviour scale of the self-report version were insufficiently reliable to warrant their empirical use. Nonetheless, criterion validity evidence was found for the conduct difficulties scale of both SDQ versions and the parent-reported social difficulties scale, suggesting that these scales are useful for screening purposes. At first sight, this seems puzzling: Reliability is traditionally regarded to be a prerequisite for validity (Moss, 1994), yet we found validity evidence for seemingly unreliable scales.

A potential explanation for the supposed lack of reliability evidence combined with evidence in favor of the scales' criterion validity can be found in what the ρ_{NL} coefficient regards to be relevant information in item scores. The ρ_{NL} coefficients in Chapter 2 were calculated based on nonlinear structural equation models. These models assume that observed item scores are comprised of a common part (i.e., what each item within a scale contributes to measuring a latent construct) and a unique part. This unique part consists of a structural component (i.e., this is potentially valuable item score content that is related to relevant external outcomes, but it is not related to what items within an

SDQ scale commonly measure) and a measurement error component (i.e., measurement inaccuracy). The ρ_{NL} coefficient expresses the extent to which all items within a scale contribute to measuring the commonly measured construct; it regards everything else, including the structural component of items, to be measurement error (Sijtsma & van der Ark, 2015).

One way to gain insight into the structural unique item components that are lost when they are not distinguished from measurement error, is by considering the content of items within a scale. As was described in the introduction to this thesis, the SDQ scales were explicitly designed with the diagnostic criteria for certain disorders in mind (Goodman, 1997; Goodman et al., 1998; Goodman & Scott, 1999) and the questionnaire has been accepted by mental healthcare professionals and researchers from all around the world (i.e., evidence for face validity). This information, combined with my own visual inspection of the conduct difficulties scale of both SDQ versions and the parent-reported social difficulties scale, suggests that the items within each of these scales seem to have something in common. Additionally, each item seems to structurally measure something unique, which makes sense as all items within a scale were intended to be different operationalizations of the same domain of psychosocial behaviour. This unique part of items is not related to what items within an SDQ scale commonly measure, yet it adds to the structural part of scale scores (i.e., in classical test theory sense: the true part). Hence, it can contribute to predicting external outcomes (i.e., belonging to a clinical population or not, and if so, the diagnosed disorder(s)) and therewith increase a scale's validity.

Further insight into the structural components of items within scales can be gained by considering the size of their unique part (i.e., standardized item residuals, not presented in this thesis). Within the conduct difficulties scale of both SDQ versions and the parent-reported social difficulties scale, the unique parts of some items were rather large, suggesting the potential presence of a substantial structural component. For example, the third item of the parent-reported conduct difficulties scale (often fights with other children) was identified as a useful indicator for the construct commonly measured by the items of the scale, yet its unique part is relatively large. A visual check of the item contents shows that this particular item is the only item within the scale that regards physical aggression. It is not unlikely that this type of behaviour is related to, for instance, belonging to a clinical population.

Note that reliability expresses measurement precision of SDQ scales for a population of adolescents (Mellenbergh, 1996; Nicewander, 2018). That is, reliability is the ratio of true SDQ scale score variance to the observed SDQ scale score variance in a population of individuals, with the observed scale scores consisting of a true score component and an error component. Reliability is only directly related to the conditional precision (i.e., measurement precision for an individual with a specific true score) of a scale score when it could be assumed that, for example, the SDQ conduct difficulties scale would measure equally precise across the full range of conduct problem severity levels in that group (i.e.,

across the full range of the latent trait measured by the scale). In reality this assumption might not hold. That is, the SDQ conduct difficulties scale might be more accurate among adolescents with severe conduct problems (i.e., limited to the high end of the range of the latent trait measured by the scale; conditional measurement precision). The conduct difficulties scale of both SDQ versions and the parent-reported social difficulties scale likely measure fairly accurately among adolescents with mid to high conduct and/or social problem severity, as these scales were found to be useful for identifying adolescents at risk of CD/ODD and ASD, respectively. Conditional measurement precision was not part of the studies in this thesis, but it can be investigated with the test information function from the Item Response Theory (IRT) framework (Samajima, 1994).

In conclusion, I deem the indications of insufficient reliability expressed in the form of ρ_{NL} coefficients on their own not enough reason to discourage the use of these scales for their intended use. In fact, given the availability of criterion validity evidence for the conduct difficulties scale of both SDQ versions and the parent-reported social difficulties scale, the information contained in scores on these scales seems useful. That is not the case for the two other scales that were found to be insufficiently reliable: the adolescent self-reported social difficulties and prosocial behaviour scales. For these scales, both reliability and validity evidence was lacking. I deem these scales of limited use for screening purposes.

Dutch SDQ norms

Chapter 6 presents gender-specific norms and joint norms (percentiles and cutoffs) per year of age for the self-report and the parent-report SDQ versions for use among Dutch adolescents aged 12 to 17. Compared to the joint norms currently available (Goodman, 1997; Goodman et al., 1998; Maurice-Stam et al., 2018), I presume our newly established norms to be more useful for interpreting Dutch adolescents' scores, because they are a) recent, b) age-specific, c) available for both SDQ versions, d) established using regression based (i.e., continuous) norming, and e) based on Dutch samples of decent size, with representativity issues corrected for.

Strengths and limitations

This thesis provides information about a broad range of validity aspects of both the self-report and parent-report SDQ versions for use among adolescents in community and clinical settings. Our samples cover a large variety in type and severity of mental health problems (including the absence of problems). I therefore presume that our samples cover the types and severity of problems found in the Dutch community and clinical populations. Here, I will discuss three limitations that are relevant with respect to each of the studies covered in this thesis, one limitation that is specifically relevant to the studies into criterion validity aspects, and one limitation that is specifically relevant to the reliability findings presented in Chapter 2.

The first limitation relevant for all studies in this thesis, concerns the representativeness of the samples. Adolescents (and parents) with another than Dutch ethnic background were substantially underrepresented in the samples used. It is unclear to what extent our findings can be generalized to ethnic minorities in the Netherlands, but findings from other countries suggest that some differences between ethnic groups are to be expected (e.g., Richter, Sagatun, Heyerdahl, Oppedal, & Røysamb, 2011). Further, the clinical sample is possibly not fully representative of all adolescents in need of help for psychiatric problems related to the constructs measured by the SDQ. That is, the sample consists of adolescents referred to mental healthcare. This means that the sample only contains adolescents that have been able to reach out for professional help. Besides, it means that the sample possibly contains relatively high-functioning adolescents, as was suggested in Chapter 4. A direct effect of the lack of representativeness of the clinical sample is that it is not possible to provide instructions on how to apply the SDQ score profile approach based on this sample. It is unknown if, and if so, how it affects the generalizability of the remaining findings regarding the use of the SDQ in diagnostic procedures.

The second limitation relates to how recent the data used in the studies in this thesis were. While the clinical data were all fairly recent, some of the community setting data were up to ten years old. As validity information can become outdated (Hubley & Zumbo, 2011), the results from some of our analyses could be somewhat affected. For all analyses regarding the community sample data, these data were combined with recent data. Hence, I expect the distorting effect of the older data, if present, to be rather limited.

The third limitation is related to assessing aspects of validity in community and clinical settings. As validity information is context-dependent, all relevant validity aspects should be researched in both settings. While we were able to investigate most validity aspects in both settings (when relevant), we investigated the SDQ's convergent and discriminant validity only in the community setting. The data needed to investigate these aspects within a clinical setting were not available. I cannot think of a reason to expect substantially different results in a clinical setting, yet by assessing these aspects in both settings we could have tested this hypothesis.

The fourth limitation relates to the criteria considered in our investigation of aspects of criterion validity. We assessed the SDQ's usefulness for identifying adolescents at risk of psychiatric disorders that are content-wise related to the SDQ by investigating its ability to discriminate between community and clinical (sub)samples. This may have led to some bias in our findings, because the community sample likely contains some adolescents with psychiatric disorders. Further, diagnoses considered in scientific studies often result from standardized procedures as these are considered reliable. In the chapters in this thesis, we considered diagnoses established by extensively trained and experienced professionals, based on a partially standardized procedure as implemented in routine clinical practice. Although these diagnoses may be less reliable, they may be more ecologically valid as these were the diagnoses that actually elicited a certain type of treatment in clinical practice.

The fifth and final limitation relates specifically to the reliability coefficients presented in Chapter 2. In that Chapter, McDonalds omega coefficients are presented. A reviewer of a previous version of a later chapter, i.e. Chapter 3, pointed out to us that the use of McDonalds omega coefficients was a suboptimal choice of coefficient: it does not express the reliability of the actual observed scale scores, whereas the nonlinear structural equation modelling reliability coefficients as reported in the final version of Chapter 3, do. The omega's were based on factor loadings that were derived from polychoric correlations obtained using the weighted least squares mean and variance adjusted (WLSMV) estimator. Consequently, omega expresses the estimated reliability of the continuous items scores that are hypothesized to underlie the observed categorical item scores. Therefore, the coefficients presented in Chapter 3 hold more practical relevance than the ones presented in Chapter 2.

Practical implications

Up to now, information on the reliability and validity of the SDQ among Dutch adolescents was scarce. The use of the SDQ within this group was mostly based on limited validity evidence that may or may not hold for Dutch adolescents. The findings in this thesis support the use of the SDQ, especially the parent-report version, in community and clinical settings. That is, our findings support the use of all self-reported and parent-reported SDQ scales, except the self-reported social difficulties and prosocial behaviour scales, and our findings show that scale scores gathered in community and clinical settings can both be interpreted using community-based norms. The latter is hugely important for the use of a screening instrument, because it implies that the meaning of SDQ scale scores is independent of whether or not an adolescent experiences problems. Then, and only then, high-scoring adolescents can be distinguished from the others based on their observed scores.

This thesis includes norms (i.e., all percentiles and cutoffs corresponding to the 80th and 90th percentiles) for both SDQ versions. Using these gender-specific norms and joint norms per year of age, healthcare professionals can compare an adolescent's scores to different reference groups: all similarly aged other adolescents or all similarly aged adolescents of the same gender. These norms are currently distributed among practitioners as part of the updated manual for the Dutch self-report and parent-report SDQ versions (Theunissen, de Wolff, Vugteveen, Timmerman, & de Bildt, 2019).

Even though we cannot provide instructions on how to apply the SDQ score profile approach at this time, the findings in this thesis clearly show that the future use of information contained in the SDQ can be optimized through the combined use of self-reported and parent-reported SDQ subscales. This profile approach corresponds to a stronger degree to the recommended use of multiple informants (American Psychiatric Association, 2013) and does more justice to the complexity of disorders and high comorbidity rates among adolescents.

Topics for future research

Validity of inferences based on scale scores can change over time and it highly depends on the context that measurements were gathered in (Hubley & Zumbo, 2011). Although this thesis provides comprehensive and recent information on validity evidence regarding the use of the SDQ in community and clinical settings, a few topics worth exploring remain.

First, certain groups of adolescents in the Netherlands could benefit from validity evidence specific to their group, as SDQ scale scores may bear different meaning within these groups. Examples of such groups are low-educated and low-literate adolescents, as they might interpret SDQ questions differently (Al-Tayyib, Rogers, Gribble, Villarroel, & Turner, 2002) and they are particularly vulnerable to mental health problems (Joffe & Black, 2012). Our samples covered a wide range of education levels, but did not include enough low-educated adolescents to separately study validity aspects for this group. Within these groups, it could additionally be worth exploring the usefulness of the teacher-reported SDQ version. Compared to younger children, adolescents spend significantly less time with each of their teachers. Therefore, I deemed the teacher-reported SDQ version to be less relevant among adolescents and focused on the adolescent-report and parent-report SDQ versions in this thesis. However, if the self-report SDQ version is of limited use among low-literate or low-educated adolescents, the teacher could be a valuable substitute informant.

As psychiatric disorders are usually defined in terms of symptoms and the distress or the impairment that they cause, a second topic that would be interesting to explore further is the usefulness of the SDQ impact scale combined with the SDQ strengths and difficulties scales for identifying adolescents at risk of psychiatric disorders. The impact scale is meant to measure chronicity, distress, and social impairment among adolescents that experience psychosocial difficulties as well as burden for others. As far as I know, the study presented in Chapter 4 of this thesis is the only available study that assessed the additive value of the impact scale among a sample consisting of only adolescents. We found that adding the impact scale to SDQ strength and difficulties scales somewhat improved the prediction of ADHD and CD/ODD among adolescents referred to mental health care. The additive value of the impact scale among adolescents with on average low problem severity (community sample), was not covered by the studies presented in this thesis.

A third aspect that is worthy of further investigation is whether reliability and validity aspects of SDQ scales are age dependent. A comparison of the findings from studies in this thesis, all conducted among an adolescent sample, to findings from studies among younger children (Stone, Otten, Engels, Vermulst, & Janssens, 2010) suggests that reliability and validity somewhat differ across these groups. For example, the parent-reported SDQ hyperactivity/inattention scale seems more strongly associated to the CBCL attention scale among adolescents ($r = .74$) than among younger children ($r = .69$), which implies stronger concurrent validity evidence for this scale within the former

group. Besides reliability and validity differences between adolescents and younger children, such differences could also exist within the adolescent group. That is, properties could differ across years of age, or even across smaller quantities of time, such as months. This age-aspect should be investigated in order to more accurately value an adolescent's SDQ scale scores at any given age.

Additionally, as mentioned earlier in this chapter, our findings suggest that some SDQ scales measure accurately across the full range of severity levels, whereas others potentially measure accurately among adolescents in a certain range of severity levels. This suspicion warrants further investigation, as it might further improve our understanding of what the SDQ scales are useful for (and among whom).

A final topic that requires further research is the SDQ profile approach suggested in this thesis. Although plenty of reliability and validity evidence was found for separate SDQ scales of the self-report and the parent-report SDQ versions, the findings in Chapter 5 suggest that the usefulness of the rich information contained in these SDQ versions can be further improved by simultaneously considering their scales. Hopefully, practitioners will soon be provided with the instructions on how to do so.

Samenvatting

SAMENVATTING

Naar schatting 15 tot 25 procent van alle adolescenten ervaart psychosociale problemen (Ormel et al., 2015). Om jongeren met dit soort problemen zo goed mogelijk te kunnen helpen is het nodig die problemen zo vroeg mogelijk op te merken. Het signaleren van problemen vindt doorgaans plaats in twee contexten. De eerste is een screeningscontext. Een voorbeeld van een dergelijke context is een periodiek gezondheidsonderzoek waarbij gepoogd wordt jongeren met psychosociale problemen te onderscheiden van de grote meerderheid zonder problemen. De tweede context is een klinische context. Een voorbeeld van een klinische context is een aanmelding bij een instelling voor kinder- en jeugdpsychiatrie. In deze context wordt gepoogd om bij jongeren, aangemeld vanwege zorgen over mogelijke problemen op psychosociaal gebied of psychiatrische stoornissen, vast te stellen welke problemen er precies spelen en hoe deze problemen te verklaren zijn. In beide contexten vindt onderzoek naar mogelijke problemen plaats op basis van informatie over het psychosociale gedrag van een jongere, bij voorkeur verstrekt door meerdere informanten (American Psychiatric Association, 2013). Eén van de instrumenten die daarbij gebruikt kan worden, is de Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997; Goodman, 1999). Dit instrument wordt veelvuldig toegepast in de Nederlandse Jeugdgezondheidszorg (JGZ) en Jeugd Geestelijke Gezondheidszorg (Jeugd GGZ). Om bruikbaar te zijn voor het signaleren en beschrijven van psychosociale problemen, moet de interpretatie van de schaalscores valide zijn (Hubley & Zumbo, 2011).

Er zijn verschillende typen validiteit te onderscheiden (Evers et al., 1988), waaronder begripsvaliditeit en criteriumvaliditeit. Begripsvaliditeit is gericht op de vraag of de schalen van een instrument meten wat ze beogen te meten; criteriumvaliditeit gaat over de mate waarin de schalen van een instrument samenhangen met relevante uitkomsten. Gezamenlijk geeft informatie over begrips- en criteriumvaliditeit van de SDQ informatie over de mate waarin SDQ-schalen, die zijn ontworpen om prosociaal gedrag en vier types problemen (emotionele problemen, gedragsproblemen, hyperactiviteit/aandachttekort en sociale problemen) te meten, bruikbaar zijn voor screening op psychosociale problemen in screeningscontext en kunnen bijdragen aan het diagnostische proces in een klinische context.

Validiteitsinformatie met betrekking tot het gebruik van de SDQ onder Nederlandse adolescenten was zeer beperkt met als gevolg dat er weinig bekend was over hoe bruikbaar de SDQ is voor het signaleren van psychosociale problemen binnen deze groep. De studies in dit proefschrift zijn gericht op het verzamelen van informatie over begrips- en criteriumvaliditeit van de zelfrapportage- en ouderversie van de SDQ voor gebruik onder 12- tot 17-jarige adolescenten in screeningscontext en klinische context in Nederland. Daarnaast zijn relatieve normen voor het interpreteren van SDQ-schaalscores gepresenteerd.

Data

In de studies in dit proefschrift is gebruik gemaakt van door de adolescent en ouder gerapporteerde SDQ data die in drie contexten zijn verzameld: een screeningscontext (*n* = 1.293), een jeugdzorgcontext (*n* = 124) en een jeugd geestelijke gezondheidszorgcontext (*n* = 4.282). In de screeningscontext is daarnaast data verzameld met de Child Behavior Checklist (CBCL; ouderrapportage; Achenbach, 1991a), de Youth Self-Report (YSR; zelfrapportage; Achenbach, 1991b) en de Intelligence Development Scales 2 (IDS-2; ouderrapportage; Grob, Hagmann-von Arx, Ruiter, Timmerman, & Visser, 2018).

Begripsvaliditeit

In dit proefschrift zijn drie aspecten van begripsvaliditeit onderzocht: in hoeverre a) de bedoelde schaalstructuur van de SDQ werd ondersteund door de data en schaalscores gelijke betekenis hebben in screeningscontext en klinische context, b) verwachte groepsverschillen tussen jongeren met en jongeren zonder psychosociale problemen werden gereflecteerd in hun SDQ-schaalscores, en c) SDQ-schalen samenhangen met schalen van andere instrumenten die dezelfde (of juist andere) constructen beogen te meten.

Schaalstructuur en meetinvariantie. De bevindingen in Hoofdstukken 2 en 3 wijzen er op dat de constructen gemeten door de vijf SDQ-schalen van beide SDQ-versies vrij goed van elkaar te onderscheiden waren, dat iedere schaal hoofdzakelijk één construct mat, en dat de meeste items van de vijf schalen bijdroegen aan het meten daarvan. De meest opvallende afwijking van de bedoelde schaalstructuur was dat de antwoorden van de jongeren werden beïnvloed door de positieve formulering van vijf van de vragen. Dat was niet het geval voor de ouderversie. Verder wijzen de bevindingen van Hoofdstuk 2 erop dat de betekenis van schaalscores van beide SDQ versies onafhankelijk is van de context (screening, klinisch) waarin de SDQ wordt ingevuld. Dit is wenselijk, omdat het betekent dat schaalscores verzameld in beide contexten kunnen worden geïnterpreteerd met dezelfde afkapwaarden (normen) en schaalscores uit beide contexten rechtstreeks met elkaar kunnen worden vergeleken.

Verwachte groepsverschillen. De bevindingen in Hoofdstuk 3 laten zien dat de verwachte groepsverschillen tussen jongeren met en jongeren zonder psychosociale problemen in hun SDQ-scores werden gereflecteerd: voor jongeren met psychosociale problemen werden hogere probleemernst en zwakkere prosociale vaardigheden gerapporteerd dan voor jongeren zonder psychosociale problemen. Dit wijst erop dat de SDQ-schalen geschikt zijn om een indicatie van probleemernst te geven.

Samenhang met andere instrumenten. In Hoofdstuk 3 werd verder duidelijk dat SDQ-schaalscores samenhangen met inhoudelijk vergelijkbare schalen van andere instrumenten en niet of nauwelijks samenhangen met inhoudelijk onvergelijkbare schalen van andere instrumenten. Dit wijst erop dat zowel de zelfrapportage- als de ouderversies van de SDQ de bedoelde constructen meten (prosociaal gedrag, emotionele problemen, gedragsproblemen, hyperactiviteit/aandachttekort en sociale problemen).

Criteriumvaliditeit

In de studies in dit proefschrift is de relatie tussen SDQ-schalen en twee externe criteria onderzocht. Het eerste criterium is de mate waarin SDQ-schalen geschikt zijn om jongeren met psychosociale problematiek van jongeren zonder deze problemen te onderscheiden in een screeningscontext. De bevindingen in Hoofdstuk 3 wijzen erop dat de totale problemenschaal (de som van de vier losse probleemschalen) van beide versies daar geschikt voor is. De zelfrapportage versie van de schaal is geschikter voor gebruik onder meisjes dan onder jongens. Dit verschil tussen meisjes en jongens lijkt vooral te wijten te zijn aan problemen met zelfrapportage over de ernst van sociale problemen door jongens. Verder bleken de emotionele problemenschaal, de gedragsproblemenschaal en de hyperactiviteit/aandachtstekortschaal van beide versies voor zowel jongens als meisjes geschikt voor het detecteren van respectievelijk angst-/stemmingsproblemen, gedragsstoornissen en ADHD. Ook bleek de door de ouder gerapporteerde sociale problemenschaal indicatief voor de aanwezigheid van autismespectrumstoornissen.

Het tweede criterium is de mate waarin SDQ-schalen geschikt zijn om in een klinische setting een eerste indicatie te geven van het type problemen dat mogelijk speelt. De bevindingen in Hoofdstuk 4 laten zien dat SDQ-schaalscores van zowel de zelfrapportage- als de ouderversie van de SDQ bruikbaar zijn voor het voorspellen van ADHD. Daarnaast is de door de ouder ingevulde SDQ bruikbaar voor het voorspellen van gedragsstoornissen en autismespectrumstoornissen. Angst-/stemmingsstoornissen bleken niet voldoende accuraat te voorspellen met behulp van de emotionele problemenschaal van beide SDQ versies.

Bovenstaande resultaten zijn gebaseerd op onderzoek naar het gebruik van aparte SDQ-schalen en informanten. In Hoofdstuk 5 wordt gerapporteerd over de bruikbaarheid van SDQ-profielen die de vijf schalen van beide SDQ versies combineren. De bevindingen zijn veelbelovend: de profielen bleken, meer dan de totale problemenschaal, geschikt voor het detecteren van psychosociale problemen in een screeningscontext. Ook bleken de profielen bruikbaar voor het verkrijgen van een eerste indicatie van het type (enkelvoudige of gecombineerde) problematiek, inclusief angst-/stemmingsproblematiek. Wel bleek dat door de ouder gerapporteerde SDQ-schalen informatiever waren dan door de jongere gerapporteerde schalen en dat de nauwkeurigheid van de profielaanpak mogelijk kan worden verbeterd door gebruik te maken van geslachtsspecifieke normen.

Betrouwbaarheid

Het onderzoeken van de betrouwbaarheid van schaalscores was geen hoofddoel van de studies in dit proefschrift, omdat algemeen wordt aangenomen dat valide schalen ook betrouwbaar zijn (betrouwbaarheid is een voorwaarde voor validiteit; e.g., Moss, 1994). Toch behoeven onze betrouwbaarheidsbevindingen, die in verschillende hoofdstukken beknopt gerapporteerd zijn, enige uitleg.

De resultaten van Hoofdstukken 2 en 3 wezen erop dat een aantal SDQ-schalen onvoldoende betrouwbaar waren: de gedragsproblemenschaal en de sociale problemenschaal van zowel de SDQ zelfrapportageversie als de ouderversie en de prosociaal gedragsschaal van de zelfrapportageversie. Voor een aantal van deze schalen (de gedragsproblemenschaal van beide SDQ versies en de sociale problemenschaal van de ouderversie) werd in Hoofdstukken 3, 4 en 5 toch gevonden dat zij samenhingen met de relevante uitkomsten. Er zijn twee plausibele verklaringen. Deze zijn reden om het gebruik van deze de drie laatstgenoemde schalen ondanks het schijnbare gebrek aan betrouwbaarheid toch aan te raden. De eerste verklaring is dat de items van de genoemde schalen mogelijk iets zinvols meten wat niet gemeenschappelijk is met de overige items van de schaal, maar wat wel bijdraagt aan de uiteindelijke score op die schaal en samenhangt met de relevante uitkomsten. Dit unieke deel van de items kon in het berekenen van de betrouwbaarheid niet van meetonnauwkeurigheid onderscheiden worden; de betrouwbaarheidscoëfficiënt drukt alleen de bijdrage van items aan het gezamenlijk gemeten construct uit.

De tweede mogelijke verklaring voor de relatief lage betrouwbaarheid in combinatie met bewijs voor criteriumvaliditeit van deze schalen is dat de schalen mogelijk nauwkeurig meten in de range van probleemernst die er toe doet (de middelmatige tot hoge ernst), maar niet onder jongeren met minder ernstige of geen problematiek. Dit kan leiden tot een schijnbaar gebrek aan betrouwbaarheid, omdat de betrouwbaarheidscoëfficiënt meetnauwkeurigheid over de gehele range van ernst uitdrukt.

Nederlandse SDQ normen

In hoofdstuk 6 worden twee typen relatieve normen voor gebruik van de SDQ zelfrapportageversie en ouderversie gepresenteerd: geslachtsspecifieke normen en normen waarbij geen onderscheid wordt gemaakt op basis van geslacht. Wij achten deze normen geschikter voor gebruik onder Nederlandse jongeren dan andere reeds beschikbare normen (Goodman, 1997; Goodman et al., 1998; Maurice-Stam et al., 2018), omdat de normen in dit proefschrift a) recent zijn, b) leeftijdsspecifiek zijn, c) beschikbaar zijn voor twee SDQ versies, d) berekend zijn met behulp van een continue normeringsprocedure, en e) gebaseerd zijn steekproeven van degelijke omvang.

Implicaties

Tot nu toe was weinig bekend over de betrouwbaarheid en validiteit van de zelfrapportage- en ouderversie van de SDQ voor gebruik onder Nederlandse adolescenten. De bevindingen in dit proefschrift laten zien dat de SDQ, met name de ouderversie, geschikt is voor gebruik in screeningscontext en klinische context. Om precies te zijn, de bevindingen onderbouwen het gebruik van alle schalen van beide SDQ versies, met uitzondering van de sociale problemenschaal en de prosociaal gedragsschaal van de zelfrapportageversie. In de toekomst kan het gebruik van beide SDQ versies verder

worden geoptimaliseerd door gebruik te maken van SDQ-profielen die de vijf schalen van beide SDQ versies combineren.

De bevindingen in dit proefschrift wijzen er verder op dat SDQ schaalscores, ongeacht de context waarin die zijn verzameld, geïnterpreteerd kunnen worden met behulp van dezelfde normen. Dit is ontzettend belangrijk voor een screeningsinstrument, want het betekent dat de betekenis van SDQ schaalscores onafhankelijk is van de context waarin het instrument is gebruikt en dat jongeren met problematiek onderscheiden kunnen worden van jongeren zonder problematiek. In dit proefschrift zijn normen gepresenteerd die kunnen worden gebruikt om de scores van een jongere te vergelijken met even oude jongeren van hetzelfde geslacht of met even oude jongeren in het algemeen. Deze normen zijn ook gepubliceerd in een nieuwe handleiding voor het gebruik van de zelfrapportage- en ouderversie van de SDQ binnen de Nederlandse jeugdgezondheidszorg (Theunissen, de Wolff, Vugteveen, Timmerman, & de Bildt, 2019).

References

Abela, J. R., & Hankin, B. L. (2008). *Handbook of depression in children and adolescents*. New York, USA: Guilford Press.

Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist. 4–18 and 1991 profile*. Burlington, VT, USA: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Youth Self Report and 1991 profile*. Burlington, VT, USA: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (2014). *DSM-oriented guide for the Achenbach system of empirically based assessment (ASEBA)*. Burlington, VT, USA: University of Vermont Research Center for Children, Youth, and Families

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. https://doi.org/10.1037/0033-2909.101.2.213

Achenbach, T., & Rescorla, L. (2001). *Manual for the ASEBA school-age forms & profiles: An integrated system of multi-informant assessment*. Burlington, VT, USA: University of Vermont.

Ackermann, K., Kirchner, M., Bernhard, A., Martinelli, A., Anomitri, C., Baker, R., . . . Gonzalez-Madruga, K. (2019). Relational aggression in adolescents with conduct disorder: Sex differences and behavioral correlates. *Journal of Abnormal Child Psychology, 47*(10), 1625-1637. https:// 10.1007/s10802-019-00541-6

Al-Tayyib, A. A., Rogers, S. M., Gribble, J. N., Villarroel, M., & Turner, C. F. (2002). Effect of low medical literacy on health survey measurements. *American Journal of Public Health, 92*(9), 1478-1480. https://doi.org/10.2105/AJPH.92.9.1478

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

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders, fourth edition: DSM-IV-TR*. Washington, DC, USA: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders, fifth edition: DSM-V*. Washington, DC, USA: American Psychiatric Association.

Arnfred, J., Svendsen, K., Rask, C., Jeppesen, P., Fensbo, L., Houmann, T., . . . Bilenberg, N. (2019). Danish norms for the strengths and difficulties questionnaire. *Danish Medical Journal, 66*(6), [A5546].

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer Et Al. *Journal of Management, 41*(6), 1561-1577. https://doi.org/10.1177/0149206315591075

Balazs, J., Miklósi, M., Keresztény, Á, Hoven, C. W., Carli, V., Wasserman, C., . . . Cosman, D. (2013). Adolescent subthreshold-depression and anxiety: Psychopathology, functional impairment and increased suicide risk. *Journal of Child Psychology and Psychiatry, 54*(6), 670-677. https://doi.org/10.1111/jcpp.12016

Becker, A., Hagenberg, N., Roessner, V., Woerner, W., & Rothenberger, A. (2004). Evaluation of the self-reported SDQ in a clinical setting: Do self-reports tell us more than ratings by adult informants? *European Child & Adolescent Psychiatry, 13*(2), ii17-ii24. https://doi.org/10.1007/s00787-004-2004-4

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. https://doi.org/10.1007/s00787-004-2003-5

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Björnsdotter, A., Enebrink, P., & Ghaderi, A. (2013). Psychometric properties of online administered parental Strengths and Difficulties Questionnaire (SDQ), and normative data based on combined online and paper-and-pencil administration. *Child and Adolescent Psychiatry and Mental Health, 7*(1), 40. https://doi.org/10.1186/1753-2000-7-40

Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): Factor structure and gender equivalence in Norwegian adolescents. *PloS One, 11*(5), e0152202. https://doi.org/10.1371/journal.pone.0152202

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis, 12*(1), 3-27. https://doi.org/10.1093/pan/mph001

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., . . . De Vet, H. C. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology, 277*(3), 826-832. https://doi.org/10.1136/bmj.h5527

Bossuyt, P. M., Reitsma, J. B., E Bruns, D., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . De Vet, H. C. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Clinical Chemistry and Laboratory Medicine, 41*(1), 68-73. https://doi.org/10.1136/bmj.326.7379.41

Brøndbo, P. H., Mathiassen, B., Martinussen, M., Heiervang, E., Eriksen, M., Moe, T. F., . . . Kvernmo, S. (2011). The Strengths and Difficulties Questionnaire as a screening instrument for Norwegian child and adolescent mental health services, application of UK scoring algorithms. *Child and Adolescent Psychiatry and Mental Health, 5,* 32. https://doi.org/10.1186/1753-2000-5-32

Buuren, S. v., & Fredriks, M. (2001). Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine, 20*(8), 1259-1277. https://doi.org/10.1002/sim.746

Cantwell, D. P., Lewinsohn, P. M., Rohde, P., & Seeley, J. R. (1997). Correspondence between adolescent report and parent report of psychiatric diagnostic data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*(5), 610-619. https://doi.org/10.1097/00004583-199705000-00011

Capron, C., Thérond, C., & Duyme, M. (2007). Psychometric properties of the French version of the self-report and teacher Strengths and Difficulties Questionnaire (SDQ). *European Journal of Psychological Assessment, 23*(2), 79-88. https://doi.org/10.1027/1015-5759.23.2.79

Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist scales for attention-deficit hyperactivity disorder: A receiver-operating characteristic analysis. *Journal of Consulting and Clinical Psychology, 62*(5), 1017-1025. https://doi.org/10.1037/0022-006X.62.5.1017

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Choi, J., Fan, W., & Hancock, G. R. (2009). A note on confidence intervals for two-group latent mean effect size measures. *Multivariate Behavioral Research, 44*(3), 396-406. https://doi.org/10.1080/00273170902938902

Christensen, D. L., Baio, J., Van Naarden Braun, K., Bilder, D., Charles, J., Constatino, J. N., . . . Yeargin-Allsopp, M. (2016). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *MMWR Surveillance Summaries, 65*(3), 1-23. http://doi.org/10.15585/mmwr.ss6503a1

Cohen, P., Cohen, J., Kasen, S., Velez, C. N., Hartmark, C., Johnson, J., . . . Streuning, E. (1993). An epidemiological study of disorders in late childhood and adolescence—I. age-and gender-specific prevalence. *Journal of Child Psychology and Psychiatry, 34*(6), 851-867. https://doi.org/10.1111/j.1469-7610.1993.tb01094.x

Constantino, J. N., & Gruber, C. P. (2005). *Social Responsiveness Scale (SRS).* Los Angeles, CA, USA: Western Psychological Services.

Costello, E. J., Copeland, W., & Angold, A. (2011). Trends in psychopathology across the adolescent years: What changes when children become adolescents, and when adolescents become adults? *Journal of Child Psychology and Psychiatry, 52*(10), 1015-1025. https://doi.org/10.1111/j.1469-7610.2011.02446.x

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*(4), 483-509. https://doi.org/10.1037/0033-2909.131.4.483

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837-845. https://doi.org/10.2307/2531595

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*(1), 8. https://doi.org/10.1186/1753-2000-2-8

Durbeej, N., Sörman, K., Selinus, E. N., Lundström, S., Lichtenstein, P., Hellner, C., & Halldner, L. (2019). Trends in childhood and adolescent internalizing symptoms: Results from Swedish population based twin cohorts. *BMC Psychology, 7*(1), 50. https://doi.org/10.1186/s40359-019-0326-8.

Dworzynski, K., Ronald, A., Bolton, P., & Happé, F. (2012). How different are girls and boys above and below the diagnostic threshold for autism spectrum disorders? *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(8), 788-797. https://doi.org/10.1016/j.jaac.2012.05.018

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association, 81*(394), 461-470. https://doi.org/10.2307/2289236

Evers, A. V. A. M., Caminada, H., Koning, R., Ter Laak, J., Van der Maesen de Sombreff, P, & Starren, J. (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen* [standards for the development and use of psychological and educational tests]. Amsterdam, the Netherlands: NIP.

Evers, A. V. A. M., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie).* Amsterdam, the Netherlands: NIP.

Evers, A. V. A. M., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*(4), 295-317. https://doi.org/10.1080/15305058.2010.518325

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*(1), 93-103. https://doi.org/10.1177/014662168701100107

Fergusson, D. M., Horwood, L. J., & Lynskey, M. T. (1993). Prevalence and comorbidity of DSM-III-R diagnoses in a birth cohort of 15 year olds. *Journal of the American Academy of Child and Adolescent Psychiatry., 32*(6), 1127-1134. https://doi.org/10.1097/00004583-199311000-00004

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine, 29*(7), 1043-1051. https://doi.org/10.1007/s00134-003-1761-8

Garrido, L. E., Barrada, J. R., Aguasvivas, J. A., Martínez-Molina, A., Arias, V. B., Golino, H. F., . . . Rojo-Moreno, L. (2018). Is small still beautiful for the Strengths and Difficulties Questionnaire? novel findings using exploratory structural equation modeling. *Assessment*. https://doi.org/10.1177/1073191118780461

Giannakopoulos, G., Tzavara, C., Dimitrakaki, C., Kolaitis, G., Rotsika, V., & Tountas, Y. (2009). The factor structure of the Strengths and Difficulties Questionnaire (SDQ) in Greek adolescents. *Annals of General Psychiatry, 8*(20), 1-7. https://doi.org/10.1186/1744-859X-8-20

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. https://doi.org/10.1007/s10802-010-9434-x

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581-586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

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*(5), 791-799. https://doi.org/10.1111/1469-7610.00494

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(11), 1337-1345. https://doi.org/10.1097/00004583-200111000-00015

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *The British Journal of Psychiatry, 177*, 534-539. https://doi.org/10.1192/bjp.177.6.534

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. https://doi.org/10.1007/s007870050057

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 & Adolescent Psychiatry, 9*(2), 129-134. https://doi.org/10.1007/s007870050008

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. https://doi.org/10.1023/a:1022658222914

Gove, W. R., & Geerken, M. R. (1977). Response bias in surveys of mental health: An empirical investigation. *American Journal of Sociology, 82*(6), 1289-1317. https://doi.org/10.1086/226466

Greene, R. W., Biederman, J., Zerwas, S., Monuteaux, M. C., Goring, J. C., & Faraone, S. V. (2002). Psychiatric comorbidity, family dysfunction, and social impairment in referred youth with oppositional defiant disorder. *American Journal of Psychiatry, 159*(7), 1214-1224. https://doi.org/10.1176/appi.ajp.159.7.1214

Grob, A., Hagmann-von Arx, P., Ruiter, S., Timmerman, M. E., & Visser, L. (2018). *IDS-2: Intelligentie- en ontwikkelingsschalen voor kinderen en jongeren*. Amsterdam, The Netherlands: Hogrefe.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Harrell, F. E. (2015). *Regression modeling stategies* (2nd ed.). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7

Harrell, F. E. (2017). Rms: Regression modeling strategies. R Package Version 5.1-4.

Hartman, C. A., Luteijn, E., Serra, M., & Minderaa, R. (2006). Refinement of the Children's Social Behavior Questionnaire (CSBQ): An instrument that describes the diverse problems seen in milder forms of PDD. *Journal of Autism and Developmental Disorders, 36*(3), 325-342. https://doi.org/10.1007/s10803-005-0072-z

He, J., Burstein, M., Schmitz, A., & Merikangas, K. R. (2013). The Strengths and Difficulties Questionnaire (SDQ): The factor structure and scale validation in US adolescents. *Journal of Abnormal Child Psychology, 41*(4), 583-595. https://doi.org/10.1007/s10802-012-9696-6

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2015). *Handbook of cluster analysis*. Boca Raton, FL, USA: CRC Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219-230. https://doi.org/10.1007/s11205-011-9843-4

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology., 3*, 29-51. https://doi.org/10.1146/annurev.clinpsy.3.022806.091419

Jansen, D. E., Wiegersma, P., Ormel, J., Verhulst, F. C., Vollebergh, W. A., & Reijneveld, S. A. (2013). Need for mental health care in adolescents and its determinants: The TRAILS Study. *The European Journal of Public Health*, *23*(2), 236-241

Jensen-Doss, A., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2014). Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *Journal of Consulting and Clinical Psychology, 82*(6), 1151-1162. https://doi.org/10.1037/a0036657

Joffe, V. L., & Black, E. (2012). Social, emotional, and behavioral functioning of secondary school students with low academic and language performance: Perspectives from students, teachers, and parents. *Language, Speech, and Hearing Services in Schools, 43*(4), 461-473. https://doi.org/10.1044/0161-1461(2012/11-0088)

Jorgensen, T., Pornprasertmanit, S., Schoemann, A., & Rosseel, Y. (2018). semTools: Useful Tools for Structural Equation Modeling. R Package Version 0.5-1.

Kattan, M. W., & Marasco, J. (2010). What is a real nomogram? *Seminars in Oncology, 37*(1), 23-26. https://doi.org/10.1053/j.seminoncol.2009.12.003

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York, USA: World Book Company.

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 & Adolescent Psychiatry, 9*(4), 271-276. https://doi.org/10.1007/s007870070030

Koskelainen, M., Sourander, A., & Vauras, M. (2001). Self-reported strengths and difficulties in a community sample of Finish adolescents. *European Child & Adolescent Psychiatry, 10*(3), 180-185. https://doi.org/10.1007/s007870170024

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (1999). *ADOS. Autism Diagnostic Observation Schedule. manual.* Los Angeles, USA: Western Psychological Services.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012). *ADOS. Autism Diagnostic Observation Schedule, second edition (ADOS-2). manual (part I): Modules 1-4.* Torrance, CA, USA: Western Psychological Services.

Lundh, L. -, Wångby-Lundh, M., & Bjärehed, J. (2008). Self-reported emotional and behavioral problems in Swedish 14 to 15-year-old adolescents: A study with the self-report version of the Strengths and Difficulties Questionnaire. *Scandinavian Journal of Psychology, 49*(6), 523-532. https://doi.org/10.1111/j.1467-9450.2008.00668.x

Mansbach-Kleinfeld, I., Apter, A., Farbstein, I., Levine, S. Z., & Poznizovsky, A. (2010). A population-based psychometric validation study of the Strengths and Difficulties Questionnaire–Hebrew version. *Frontiers in Psychiatry, 1*(151) 1-12. https://doi.org/10.3389/fpsyt.2010.00151

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85-110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Martel, M. M., Markon, K., & Smith, G. T. (2017). Research review: Multi-informant integration in child and adolescent psychopathology diagnosis. *Journal of Child Psychology and Psychiatry, 58*(2), 116-128. https://doi.org/10.1111/jcpp.12611

Maurice-Stam, H., Haverman, L., Splinter, A., van Oers, H., Schepers, S., & Grootenhuis, M. (2018). Dutch norms for the Strengths and Difficulties Questionnaire (SDQ)–parent form for children aged 2–18 years. *Health and Quality of Life Outcomes, 16*(1), 123. https://doi.org/10.1186/s12955-018-0948-1

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ, USA: Erlbaum.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293-299. https://doi.org/10.1037/1082-989X.1.3.293

Mellor, D. (2005). Normative data for the Strengths and Difficulties Questionnaire in Australia. *Australian Psychologist, 40*(3), 215-222. https://doi.org/10.1080/00050060500243475

Merikangas, K. R., He, J., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., . . . Swendsen, J. (2010). Lifetime prevalence of mental disorders in US adolescents: Results from the National Comorbidity Survey replication–Adolescent supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry, 49*(10), 980-989. https://doi.org/10.1016/j.jaac.2010.05.017

Messick, S. (1989). Validity. In. R. L. Linn (ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York, NY, USA: Macmillan Publishing.

Metz, C. E. (1978). Basic principles of ROC analysis. Paper presented at the *Seminars in Nuclear Medicine, 8*(4) 283-298. https://doi.org/10.1016/s0001-2998(78)80014-2

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515. https://doi.org/10.1207/S15327906MBR3903_4

Moriwaki, A., & Kamio, Y. (2014). Normative data and psychometric properties of the Strengths and Difficulties Questionnaire among Japanese school-aged children. *Child and Adolescent Psychiatry and Mental Health, 8*(1), 1-12. https://doi.org/10.1186/1753-2000-8-1

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12. https:/doi.org/10.2307/1176218

Mowlem, F., Agnew-Blais, J., Taylor, E., & Asherson, P. (2019). Do different factors influence whether girls versus boys meet ADHD diagnostic criteria? Sex differences among children with high ADHD symptoms. *Psychiatry Research, 272*, 765-773. https://doi.org/10.1016/j.psychres.2018.12.128

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132. https://doi.org/10.1007/BF02294210

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (Eight Edition ed.). Los Angeles, CA, USA: Muthén & Muthén.

Nakamura, B. J., Ebesutani, C., Bernstein, A., & Chorpita, B. F. (2009). A psychometric analysis of the Child Behavior Checklist DSM-oriented scales. *Journal of Psychopathology and Behavioral Assessment, 31*(3), 178-189. https://doi.org/10.1007/s10862-008-9119-8

Nanninga, M., Jansen, D. E., Knorth, E. J., & Reijneveld, S. A. (2018). Enrolment of children in psychosocial care: problems upon entry, care received, and outcomes achieved. *European Child Adolescent Psychiatry, 27*(5), 625-635.

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods, 23*(2), 351-362. https://doi.org/10.1037/met0000132

Olfson, M., Blanco, C., Wang, S., Laje, G., & Correll, C. U. (2014). National trends in the mental health care of children, adolescents, and adults by office-based physicians. *JAMA Psychiatry, 71*(1), 81-90. https://doi.org/10.1001/jamapsychiatry.2013.3074

Olfson, M., Druss, B. G., & Marcus, S. C. (2015). Trends in mental health care among children and adolescents. *The New England Journal of Medicine, 372*, 2029-2038. https://doi.org/10.1056/NEJMsa1413512

Ormel, J., Raven, D., van Oort, F., Hartman, C., Reijneveld, S., Veenstra, R., . . . Oldehinkel, A. (2015). Mental health in Dutch adolescents: A TRAILS report on prevalence, severity, age of onset, continuity and co-morbidity of DSM disorders. *Psychological Medicine, 45*(02), 345-360. https://doi.org/10.1017/S0033291714001469

Ortuño-Sierra, J., Fonseca-Pedrero, E., Paino, M., Sastre i Riba, S., & Muñiz, J. (2015). Screening mental health problems during adolescence: Psychometric properties of the Spanish version of the Strengths and Difficulties Questionnaire. *Journal of Adolescence, 38*, 49-56. https://doi.org/10.1016/j.adolescence.2014.11.001

Phillips, D. L., & Clancy, K. J. (1970). Response biases in field studies of mental illness. *American Sociological Review, 35*(3), 503-515. https://doi.org/10.2307/2092992

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*(3), 603-610. https://doi.org/10.1177/0013164490503016

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Raiker, J. S., Freeman, A. J., Perez-Algorta, G., Frazier, T. W., Findling, R. L., & Youngstrom, E. A. (2017). Accuracy of Achenbach scales in the screening of attention-deficit/hyperactivity disorder in a community mental health clinic. *Journal of the American Academy of Child & Adolescent Psychiatry, 56*(5), 401-409. https://doi.org/10.1016/j.jaac.2017.02.007

Reijneveld, S. A., Wiegersma, P. A., Ormel, J., Verhulst, F. C., Vollebergh, W. A., & Jansen, D. E. (2014). Adolescents' use of care for behavioral and emotional problems: Types, trends, and determinants. *PloS One, 9*(4), e93526. https://doi.org/10.1371/journal.pone.0093526

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*(3), 169-184. https://doi.org/10.1002/mpr.289

Rey, J. M., Schrader, E., & Morris-Yates, A. (1992). Parent-child agreement on children's behaviours reported by the Child Behaviour Checklist (CBCL). *Journal of Adolescence, 15*(3), 219-230. https://doi.org/10.1016/0140-1971(92)90026-2

Richter, J., Sagatun, Å, Heyerdahl, S., Oppedal, B., & Røysamb, E. (2011). The Strengths and Difficulties Questionnaire (SDQ)–Self-Report. an analysis of its structure in a multi-ethnic urban adolescent sample. *Journal of Child Psychology and Psychiatry, 52*(9), 1002-1011. https://doi.org/10.1111/j.1469-7610.2011.02372.x

Rigby, R. A., & Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution. *Statistics in Medicine, 23*(19), 3053-3076. https://doi.org/10.1002/sim.1861

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(3), 507-554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. -., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(77), 1-8. https://doi.org/10.1186/1471-2105-12-77

Rønning, J. A., Handegaard, B. H., Sourander, A., & Mørch, W. (2004). The Strengths and Difficulties self-report Questionnaire as a screening instrument in Norwegian community samples. *European Child & Adolescent Psychiatry, 13*(2), 73-82. https://doi.org/10.1007/s00787-004-0356-4

Ruchkin, V., Koposov, R., & Schwab-Stone, M. (2007). The strength and difficulties questionnaire: Scale validation with Russian adolescents. *Journal of Clinical Psychology, 63*(9), 861-869. https://doi.org/10.1002/jclp.20401

Russell, G., Rodgers, L. R., & Ford, T. (2013). The Strengths and Difficulties Questionnaire as a predictor of parent-reported diagnosis of autism spectrum disorder and attention deficit hyperactivity disorder. *PloS One, 8*(12), e80247. https://doi.org/10.1371/journal.pone.0080247

Salbach-Andrae, H., Lenz, K., & Lehmkuhl, U. (2009). Patterns of agreement among parent, teacher and youth ratings in a referred sample. *European Psychiatry, 24*(5), 345-351. https://doi.org/10.1016/j.eurpsy.2008.07.008

Samajima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*(3), 229-244. https://doi.org/10.1177/014662169401800304

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561-582. https://doi.org/10.1080/10705510903203433

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity, 24*(4), 367-386. https://doi.org/10.1007/BF00152011

Schachar, R., & Wachsmuth, R. (1989). *Parent Interview for Child Symptoms – revised DSM-III-R*. Toronto, Canada: Hospital for Sick Children, Department of Psychiatry.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114. https://doi.org/10.1177/001316448104100420

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Sijtsma, K., & van der Ark, L A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research, 64*(2), 128-136. https://doi.org/10.1097/NNR.0000000000000077

Silverman, W. K., & Albano, A. M. (1996). *The anxiety disorders interview schedule for children for DSM-IV: (Child and parent versions)*. San Antonio, TX, USA: Psychological Corporation.

Smith, S. R. (2007). Making sense of multiple informants in child and adolescent psychopathology: A guide for clinicians. *Journal of Psychoeducational Assessment, 25*(2), 139-149. https://doi.org/10.1177/0734282906296233

Smits, I. A., Theunissen, M. H. C., Reijneveld, S. A., Nauta, M. H., & Timmerman, M. E. (2016). Measurement invariance of the parent version of the Strengths and Difficulties Questionnaire (SDQ) across community and clinical populations. *European Journal of Psychological Assessment, 34*(4), 238-246. https://doi.org/10.1027/1015-5759/a000339

Statistics Netherlands. (2015). Statline. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37296ned/table?ts=152209294

Steiger, J. H. (1980). Statistically based tests for the number of common factors. Paper presented at the *Paper Presented at the Annual Meeting of the Psychometric Society, Iowa City, IA, USA, may 1980,*

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. https://doi.org/10.1007/s10567-010-0071-2

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293. https://doi.org/10.1126/science.3287615

Tannock, R., Hum, M., Humphries, T., & Schachar, R. (2000). Interviewing teachers about children's classroom behavior and academic performance. Paper presented at the *Proceedings of the 47th Annual Meeting of the American Academy of Child and Adolescent Psychiatry,* New York, NY, USA, October 24-29.

Theunissen, M. H. C., de Wolff, M. S., & Reijneveld, S. A. (2019). The Strengths and Difficulties Questionnaire self-report: A valid instrument for the identification of emotional and behavioral problems. *Academic Pediatrics, 19*(4), 471-476. https://doi.org/10.1016/j.acap.2018.12.008

Theunissen, M. H. C., de Wolff, M. S., Van Grieken, A., & Mieloo, C. (2016). *Handleiding voor het gebruik van de SDQ binnen de jeugdgezondheidszorg. vragenlijst voor het signalering van psychosociale problemen bij 3-17 jarigen.* Leiden, the Netherlands: TNO.

Theunissen, M. H. C., de Wolff, M. S., Vugteveen, J., Timmerman, M. E., & de Bildt, A. (2019). *Handleiding voor het gebruik van de SDQ bij adolescenten (12-17 jaar) binnen de jeugdgezondheidszorg. vragenlijst voor het signaleren van psychosociale problemen*. Leiden, the Netherlands: TNO.

Timmerman, M. E., Voncken, L., & Albers, C. A. (2019). A tutorial on regression-based norming of psychological tests with GAMLSS. https://doi.org/10.31219/osf.io/mdc9u

Tobia, V., & Marzocchi, G. M. (2018). The Strengths and Difficulties Questionnaire-parents for Italian school-aged children: Psychometric properties and norms. *Child Psychiatry & Human Development, 49*(1), 1-8. https://doi.org/10.1007/s10578-017-0723-2

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. https://doi.org/10.1007/BF02291170

van de Looij-Jansen, P M, Goedhart, A. W., de Wilde, E. J., & Treffers, P. D. A. (2011). Confirmatory factor analysis and factorial invariance analysis of the adolescent self-report Strengths and Difficulties Questionnaire: How important are method effects and minor factors? *British Journal of Clinical Psychology, 50*(2), 127-144.

Van der Ark, L A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19. https://doi.org/10.18637/jss.v020.i11

van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology, 60*(2), 315-337. https://doi.org/10.1348/000711006X117574

van Lang, N. D., Ferdinand, R. F., Oldehinkel, A. J., Ormel, J., & Verhulst, F. C. (2005). Concurrent validity of the DSM-IV scales affective problems and anxiety problems of the youth self-report. *Behaviour Research and Therapy, 43*(11), 1485-1494. https://doi.org/10.1016/j.brat.2004.11.005

van Roy, B., Veenstra, M., & Clench-Aas, J. (2008). Construct validity of the five-factor Strengths and Difficulties Questionnaire (SDQ) in pre-, early, and late adolescence. *Journal of Child Psychology and Psychiatry, 49*(12), 1304-1312. https://doi.org/10.1111/j.1469-7610.2008.01942.x

Van Widenfelt, B. M., Goedhart, A. W., Treffers, P. D. A., & Goodman, R. (2003). Dutch version of the Strengths and Difficulties Questionnaire (SDQ). *European Child & Adolescent Psychiatry, 12*(6), 281-289. https://doi.org/10.1007/s00787-003-0341-3

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymetry (SOKA) model. *Journal of Personality and Social Psychology, 98*(2), 281-300. https://doi.org/10.1037/a0017908

Verhulst, F. C., Van der Ende, J., & Koot, H. M. (1996). *Dutch manual for the CBCL/4–18*. Rotterdam, the Netherlands: Afdeling Kinder- en Jeugdpsychiatrie, Sophia Kinderziekenhuis / Academisch Ziekenhuis Rotterdam, Erasmus Universiteit Rotterdam.

Verhulst, F. C., Van der Ende, J., & Koot, H. M. (1997). *Dutch manual for the youth self-report (YSR)*. Rotterdam, the Netherlands: Afdeling Kinder- en Jeugdpsychiatrie, Sophia Kinderziekenhuis / Academisch Ziekenhuis Rotterdam, Erasmus Universiteit Rotterdam.

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.5. User's Guide*. Belmont, MA, USA: Statistical Innovations Inc.

Vogels, A. G. C., Siebelink, B. M., Theunissen, M. H. C., de Wolff, M. S., & Reijneveld, S. A. (2011). *Vergelijking van de KIVPA en de SDQ als signaleringsinstrument voor problemen bij adolescenten in de jeugdgezondheidszorg* Leiden, the Netherlands: TNO.

Vollebergh, W. A., van Dorsselaer, S., Monshouwer, K., Verdurmen, J., van der Ende, J., & ter Bogt, T. (2006). Mental health problems in early adolescents in the Netherlands. *Social Psychiatry and Psychiatric Epidemiology, 41*(2), 156-163. https://doi.org/10.1007/s00127-005-0979-x

Vugteveen, J., De Bildt, A., Hartman, C., & Timmerman, M. E. (2018). Using the Dutch multi-informant Strengths and Difficulties Questionnaire (SDQ) to predict adolescent psychiatric diagnoses. *European Child & Adolescent Psychiatry, 27*, 1347-1359. https://doi.org/10.1007/s00787-018-1127-y

Vugteveen, J., de Bildt, A., Serra, M., de Wolff, M. S., & Timmerman, M. E. (2018). Psychometric properties of the Dutch Strengths and Difficulties Questionnaire (SDQ) in adolescent community and clinical populations. *Assessment,* https://doi.org/10.1177/1073191118804082

Vugteveen, J., de Bildt, A., Theunissen, M. H. C., Reijneveld, S. A., & Timmerman, M. E. (2019). Validity aspects of the Strengths and Difficulties Questionnaire (SDQ) adolescent self-report and parent-report versions among Dutch adolescents. *Assessment,* https://doi.org/10.1177/1073191119858416

Waschbusch, D. A., King, S., & Nothern Partners in Action for Children and Youth (2006). Should sex-specific norms be used to assess attention-deficit/hyperactivity disorder or oppositional defiant disorder? *Journal of Consulting and Clinical Psychology, 74*(1), 179-185. https://doi.org/10.1037/0022-006X.74.1.179

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology* (pp. 50-81). Hoboken, NJ, USA: John Wiley & Sons. https://doi.org/10.1002/9781118133880.hop210003

Woerner, W., Nuanmanee, S., Becker, A., Wongpiromsarn, Y., & Mongkol, A. (2011). Normative data and psychometric properties of the Thai version of the Strengths and Difficulties Questionnaire (SDQ). *Journal of Mental Health of Thailand, 19*(1), 42-57.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* Geneva, Switzerland: World Health Organization.

Yang, Y., & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology, 11*(1), 23–34. https://doi.org/10.1027/1614-2241/a000087

Youngstrom, E. (2013). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child & Adolescent Psychology, 42*(1), 139-159. https://doi.org/10.1080/15374416.2012.736358

Youngstrom, E., & Frazier, T. (2013). Evidence-based strategies for the assessment of children and adolescents: Measuring prediction, prescription, and process. In D. J. Miklowitz, W. E. Craighead, & L. Craighead (Eds.), *Developmental psychopathology* (2nd ed., pp. 36–79). New York, NY, USA: Wiley.

Dankwoord

Curriculum vitae

List of publications

DANKWOORD

Marieke en Annelies, wat vormden we een mooi team. Marieke, jouw gedrevenheid, gestructureerde manier van denken en methodologische kennis waren inspirerend en hielpen mij het onderste uit de kan te halen. Annelies, jouw inhoudelijke kennis over screening en diagnostiek hielp me de zorgpraktijk beter te begrijpen, met als resultaat dat mijn werk zowel wetenschappelijk als maatschappelijk relevant is. Onze overleggen waren gezellig, enerverend en motiverend. Mede dankzij jullie is dit proefschrift tot stand gekomen. Ik ben er trots op. Heel erg bedankt voor jullie begeleiding!

Selma, bedankt voor jouw grote bijdrage aan de dataverzameling. Ik waardeer de onophoudelijke inzet voor het werven van participanten en de vindingrijkheid waarmee je het voor hen zo aantrekkelijk mogelijk maakte om vragenlijsten voor ons in te vullen.

Marianne en Meinou, ik wil jullie bovenal bedanken voor jullie bijdrage aan het vertalen van mijn onderzoeksresultaten naar concrete adviezen voor het gebruik van de Strengths and Difficulties Questionnaire in de praktijk. Ik hecht veel waarde aan deze vertaalslag.

Catharina, Marianne, Marike, Meinou en Menno, heel erg bedankt voor jullie bijdragen aan de artikelen die de basis vormen voor de hoofdstukken in dit proefschrift. Jullie kritische en constructieve blik heeft bijgedragen aan werk waar ik trots op ben.

Daniela and Lieke, we shared three year of ups and downs in our personal and professional lives. We laughed, we cried, and we laughed some more. Thanks for your support and the gezelligheid.

To all my other colleagues, from the Psychometrics and Statistics group and beyond, thank you for your support and company. A special thank you goes to Jorge. Jorge, sharing an office with you in the final stages of my PhD project was awesome. Thank you for your kindness and positivity!

Jacco en Rink, hoewel jullie niet direct bij mijn promotieonderzoek betrokken zijn geweest, wil ik jullie hier toch noemen omdat mijn interesse in statistiek en de wetenschap zeker ook aan jullie te danken is. Rink, bedankt voor jouw betrokkenheid en vriendschap. Ik denk nog steeds met weemoed terug aan onze fijne samenwerking tijdens mijn GION-jaren. Jacco, jouw kritische reflectie op alles wat me maar bezig hield gedurende de afgelopen tien jaar en de support en adviezen die daarmee gepaard gingen, betekenen meer voor me dan ik hier in woorden uit kan drukken. Dankjewel.

Bas en Linda, ik kan altijd bij jullie terecht voor opbeurende woorden, een kritische blik of gewoon om te kletsen over van alles en nog wat. Ik waardeer jullie vriendschap enorm. Bedankt dat jullie mijn paranimfen willen zijn!

Jorien Vugteveen
November, 2020

CURRICULUM VITAE

Jorien Vugteveen completed a teacher training program for teachers in primary education at CHN / Stenden University of Applied Science in Emmen. After obtaining her teaching degree in 2010, she moved on to the master's program in Educational sciences at the University of Groningen. In 2012, she completed the program with a specialisation in educational psychology and school consultation, and a specialisation in curriculum and instruction.

In 2012, Jorien was appointed at the Department of Educational and Pedagogical sciences of the University of Groningen were she taught research methodology / statistics courses, and was involved in several research projects. In 2016, she started her PhD research on the measurement quality of the Strengths and Difficulties Questionnaire for assessing psychosocial behaviour among Dutch adolescents. During this PhD project at the Department of Psychology (Psychometrics and Statistics) at the University of Groningen, she was supervised by Prof. dr. Marieke Timmerman and Dr. Annelies de Bildt. The project was funded by ZonMw – the Netherlands Organization for Health Research and Development. Jorien presented at several national and international conferences, including the International Test Commission Conference (Montréal, 2018) where she was awarded the Best Student Paper Award for the paper that Chapter 4 of this thesis is based on. Her research findings are published in scientific journals as well as journals aimed at reaching professionals in Dutch social care and mental healthcare (In Dutch: Jeugdgezondheidszorg (JGZ) and Jeugd Geestelijke Gezondheidszorg (Jeugd GGZ)). As of May 2019, she works as an assistant professor at the Department of Psychology (Psychometrics and Statistics) at the University of Groningen.

LIST OF PUBLICATIONS

Peer-reviewed publications

Published

Vugteveen, J., de Bildt, A., Hartman, C., Reijneveld, S.A ., & Timmerman, M. E. (Accepted). The combined self- and parent-rated SDQ score profile predicts care use and psychiatric diagnoses. *European Child & Adolescent Psychiatry*.

Vugteveen, J., de Bildt, A., Theunissen, M.S., Reijneveld, S.A., & Timmerman, M.E. (2019). Validity aspects of the Strengths and Difficulties Questionnaire (SDQ) adolescent self-report and parent versions among Dutch adolescents. *Assessment*. https://doi.org/10.1177/1073191119858416

Hoekstra, R., **Vugteveen, J.**, Warrens, M.J., & Kruyen, P.M. (2019*).* An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology, 22*(4), 351-364. https://doi.org/10.1080/13645579.2018.1547523

Elgersma, H.J., Koster, E.H.W., **Vugteveen, J.**, Hoekzema, A., Penninx, B.W.J.H., Bockting, C.L.H., & de Jong, P.J. (2019). Predictive Value of Attentional Bias for the Recurrence of Depression: A 4-year Prospective Study in Remitted Depressed Individuals. *Behaviour Research and Therapy, 114*, 25-34. https://doi.org/10.1016/j.brat.2019.01.001

Vugteveen, J., Boevé, A., & Hoekstra, R. (2018). The struggles of a field study: Studying the effectiveness of a flipped classroom. *SAGE Research Methods Cases*, (2). https://doi.org/10.4135/9781526438188

Vugteveen, J., de Bildt, A., Hartman, C. A., & Timmerman, M.E. (2018). Using the Dutch multiinformant Strengths and Difficulties Questionnaire (SDQ) to predict adolescent psychiatric diagnoses. *European Child & Adolescent Psychiatry*, 27(10), 1347-1359. https://doi.org/10.1007/s00787-018-1127-y

Vugteveen, J., de Bildt, A., Serra, M., de Wolff, M.S., & Timmerman, M.E. (2018). Psychometric Properties of the Dutch Strengths and Difficulties Questionnaire (SDQ) in Adolescent Community and Clinical Populations. *Assessment*. https://doi.org/10.1177/1073191118804082

Boevé, A. J., Meijer, R. R., Bosker, R. J., **Vugteveen, J.**, Hoekstra, R., & Albers, C. J. (2017). Implementing the Flipped Classroom: An exploration of study behaviour and student performance. *Higher Education*, 74(6), 1015-1032. https://doi.org/10.1007/s10734-016-0104-y

van Rooijen, M., Korpershoek, H., **Vugteveen, J.**, & Opdenakker, M. -C. (2017). De overgang van het basis- naar het voortgezet onderwijs en de verdere schoolloopbaan. *Pedagogische studien*, 94(2), 110-134.

Vugteveen, J., & Timmermans, A. (2017). Risico op afbreken van een opleiding in de eerste twee jaar van het mbo. *Pedagogische Studiën*, 94(3), 139-159.

Submitted

Vugteveen, J., de Bildt, A., & Timmerman, M.E. Dutch normative data for the self-reported and parent-reported Strengths and Difficulties Questionnaire (SDQ) for ages 12-17.

von Spreckelsen, P., Jonker, N., **Vugteveen, J**., Wessel, I., Glashouwer, K. A., & de Jong, P. J. Individual Differences in Avoiding Feelings of Disgust: Development and Construct Validity of the Disgust Avoidance Questionnaire. (preprint: https://doi.org/10.31234/osf.io/sw674)

Professional publications

Published

Theunissen, M., de Wolff, M., **Vugteveen, J.**, Timmerman, M. E., & de Bildt, A. (2019). *Handleiding voor het gebruik van de Strengths and Difficulties Questionnaire bij adolescenten (12-17 jaar) binnen de Jeugdgezondheidszorg: Vragenlijst voor het signaleren van psychosociale problemen*. Leiden: TNO.

Vugteveen, J., Timmerman, M., de Bildt, A., & de Wolff, M. (2017). Hoe goed signaleert de SDQ problemen bij jongeren? *Kind en Adolescent Praktijk*, 16(3), 34-36. https://doi.org/10.1007/s12454-017-0033-7

Korpershoek, H., Beijer, C., Spithoff, M., Naaijer, H. M., Timmermans, A. C., van Rooijen, M., **Vugteveen, J.**, & Opdenakker, M-C. (2016). *Overgangen en aansluitingen in het onderwijs: Deelrapportage 1: reviewstudie naar de po-vo en de vmbo-mbo overgang*. Groningen: GION onderzoek/onderwijs.

van Rooijen, M., Korpershoek, H., **Vugteveen, J.**, Timmermans, A. C., & Opdenakker, M-C. (2016). *Overgangen en aansluitingen in het onderwijs: Deelrapportage 2: empirische studie naar de cognitieve en niet-cognitieve ontwikkeling van leerlingen rondom de po-vo overgang*. GION onderzoek/onderwijs.

Vugteveen, J., Timmermans, A. C., Korpershoek, H., van Rooijen, M., & Opdenakker, M-C. (2016). *Overgangen en aansluitingen in het onderwijs: Deelrapportage 3: empirische studie naar de cognitieve en niet-cognitieve ontwikkeling van leerlingen rondom de vmbo-mbo overgang*. Groningen: GION onderzoek/onderwijs.

Vugteveen, J., & Timmermans, A. (2016). Motivatie in het mbo. *Didactief*, 46(8), 26-27.

Warrens, M. J., de Raadt, A., **Vugteveen, J.**, van Rijn, N., Korpershoek, H., Guldemond, H., ... Opdenakker, M-C. (2016). *Overgangen en aansluitingen in het onderwijs: Deelrapportage 4: draaien aan de knoppen. Een simulatiestudie naar de effecten van enkele beleidsparameters op de aansluiting po-vo.* Groningen: GION onderwijs/onderzoek.

Vugteveen, J., van der Putten, A. A. J., & Vlaskamp, C. (2014). Inventarisatieonderzoek Mensen met Ernstige Meervoudige Beperkingen: Prevalentie en Karakteristieken. Groningen: Stichting Kinderstudies.

Submitted

Vugteveen, J., Timmerman, M., de Bildt, A., & de Wolff, M. Zo goed signaleert de SDQ problemen bij jongeren. *Kind en Adolescent Praktijk*