Psychometric Properties of the Dutch Strengths and Difficulties Questionnaire (SDQ) in Adolescent Community and Clinical Populations

Jorien Vugteveen[1] [iD], Annelies de Bildt[1,2] [iD], Marike Serra[1], Marianne S. de Wolff[3], and Marieke E. Timmerman[1]

Abstract

This study assessed the factor structures of the Strengths and Difficulties Questionnaire (SDQ) adolescent and parent versions and their measurement invariance across settings in clinical ($n = 4,053$) and community ($n = 962$) samples of Dutch adolescents aged 12 to 17 years. Per SDQ version, confirmatory factor analyses were performed to assess its factor structure in clinical and community settings and to test for measurement invariance across these settings. The results suggest measurement invariance of the presumed five-factor structure for the parent version and a six-factor structure for the adolescent version. Furthermore, 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 favorable but less easily computed factor scores. These findings suggest that adolescent- 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.

Keywords

Strengths and Difficulties Questionnaire, adolescent and parent versions, clinical and community settings, factor structure, measurement invariance

The Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) aims at measuring psychosocial functioning among children and adolescents aged 4 to 17 years. This widely used questionnaire is valued for three reasons. First, with only 25 items, the SDQ is relatively short. Second, the SDQ not only covers deficits (hyperactivity/inattention, conduct problems, emotional problems, peer problems) but also strengths (prosocial behavior). Third, the availability of multiple informant versions allows an individual's psychosocial functioning to be assessed from multiple perspectives. For adolescents aged 11 to 16 years, an adolescent version (also known as the self-report version) and a parent 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 (e.g., 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 SDQ parent and adolescent versions (e.g., 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, 2001) of both the SDQ adolescent and the SDQ parent version has repeatedly been

[1]University of Groningen, Groningen, Netherlands
[2]University Medical Center Groningen (UMCG), Groningen, Netherlands
[3]Netherlands Organisation of Applied Scientific Research (TNO), Leiden, Netherlands

Corresponding Author:
Jorien Vugteveen, Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, Groningen 9700 AB, Netherlands.
Email: j.vugteveen@rug.nl

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 *adolescent* 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 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 years, but some studies covered that whole age range, while others only covered 2 or 3 years of age (e.g., 14-15 or 16-18 years). The samples further differed in country of origin; most of the studies mentioned were performed in Northeast 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.

For the SDQ *parent* 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, respectively) 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.

Considering the somewhat mixed results on the tenability of the five-factor structure regarding the SDQ adolescent 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 (1997), and an additional *positive construal method* factor. The latter is composed 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 (e.g., Pilotte & Gable, 1990; Schriesheim & Hill, 1981). The positive method factor thus expresses the method effect bias resulting from combining positively and negatively worded items in the SDQ problem 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 nonclinical 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 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, Theunissen, Reijneveld, Nauta, and Timmerman (2016) found evidence for measurement invariance across these populations for the five-factor SDQ parent version among 2- to 14-year-olds. To the best of our knowledge, measurement invariance across these settings among adolescents has not been investigated.

The aim of the current study is to assess the presumed five-factor structure of the SDQ adolescent and the parent versions, and to examine their measurement invariance across community and clinical populations of Dutch adolescents aged 12 to 17 years. In case the presumed five-factor structure does not fit adequately, we will investigate 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* 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). Furthermore, we hypothesize to find measurement invariance of the five-factor SDQ parent version across the two populations, consistent with findings by Smits et al. (2016), thereby assuming that the parent's manner of judgment regarding an adolescent's psychosocial functioning does not substantially differ from their manner of judgment 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 reassurance for this sum score method.

For the SDQ *adolescent* version, we cautiously expect to find confirmation for the presumed five-factor structure as

findings from previous research regarding factor structure in community populations are mixed. With regard to the factor structure of the *adolescent* 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.

Method

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 were 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, adolescent-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 1 presents additional demographic and geographic characteristics of the clinical sample.

Table 2 provides an overview of the *Diagnostic and Statistical Manual of Mental Disorders–Fourth edition* (*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 contentwise 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 ($n = 609$, 15.1%). The second column of the table shows that anxiety/mood disorders were most prevalent, and conduct/oppositional defiant disorder least prevalent. Per *DSM-IV* disorder (row), columns three through six provide information about the comorbidity of disorders. Most prevalent is attention-deficit/hyperactivity disorder within the group with conduct/oppositional defiant disorder.

Community Sample. Within the community sample of 12- to 17-year-old adolescents, data were collected in three waves. The first wave of adolescent- 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

Table 1. Demographic and Geographic Characteristics of the Adolescents in the Clinical and Community Sample.

Characteristics	Clinical, *n* (%)	Community, *n* (%)
Gender		
Male	1,902 (46.9)[a]	474 (49.3)[b]
Female	2,093 (51.6)	482 (50.1)
Native country mother		
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, years		
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)

[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%).

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 adolescent-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 adolescent-reported SDQ data ($n = 217$), parent-reported SDQ data ($n = 28$), or both ($n = 274$) were available. The third wave of data consisted of adolescent- 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 adolescent-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 adolescent-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

Table 2. Prevalence of *DSM-IV* Diagnoses and Comorbidity Between *DSM-IV* Diagnoses.

		Comorbid with . . .			
DSM category	N^a	ADHD[b]	CD/ODD[b]	Anxiety/mood disorder[b]	ASD[b]
ADHD	913	—	.18	.14	.16
Anxiety/mood disorder	1,372	.09	.03	—	.09
ASD	719	.20	.04	.18	—
CD/ODD	391	.42	—	.09	.08

Note. DSM-IV = Diagnostic and Statistical Manual of Mental Disorders–Fourth edition; ADHD = attention-deficit/hyperactivity disorder; CD/ODD = conduct/oppositional defiant disorder; ASD = autism spectrum disorder.
[a]The numbers in this column add up to more than 2,412 (number of adolescent in the sample with a diagnosis in any of the four categories) due to comorbidity.
[b]The proportion of adolescents within each *DSM* category (row), also diagnosed with any of the other disorders.

(*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 1. When compared with 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 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 years.

Strengths and Difficulties Questionnaire

Adolescents and their parents completed the Dutch version of the SDQ adolescent and parent 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 behavior, emotional functioning, hyperactivity and interaction with peers, and one subscale of five items focusing on prosocial behavior, which is considered a strength (Goodman, 1997). For each item, a 3-point rating scale (0 = *not true*, 1 = *somewhat true*, and 2 = *certainly true*) rates the degree to which either the adolescent considers the attribute applicable to oneself, or the parent considers it 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 scale represents a high degree of prosocial behavior. 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 article.

Statistical Analysis

Missing Data. The clinical sample contained no missing data; the community sample data set contained some missing data for the SDQ adolescent version (*M* = 0.33%, *SD* = 0.32, minimum = 0%, maximum = 1.2%) and the SDQ parent version (*M* = 0.38%, *SD* = 0.28, minimum = 0%, maximum = 0.8%). Considering the small number of missing data, we opted for two-way imputation with normally distributed errors to impute these data (e.g., 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 (B. 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 (see, e.g., 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 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 nonzero loadings) holds across both the clinical and community settings. For identification of the model, the following constraints were applied (following 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 of those indices with a value larger than 10, and the model was rerun. We repeated this process until the model fitted sufficiently or the model was rerun 10 times. We chose 10 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 that 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 M*plus* version 8 (L. K. Muthén & Muthén, 1998-2017), using weighted least squares mean and variance adjusted estimation. The goodness-of-fit of the models was assessed by considering the root mean square error of approximation (RMSEA) value (Steiger, 1980) and the comparative fit index (CFI;: Bentler, 1990). We consider RMSEA values ⩽.08 combined with CFI values ⩾.90 to be acceptable, while we prefer RMSEA values ⩽.06 together with CFI values ⩾.95 are preferred, as is recommended by Hu and 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, considering a decrease of .01 or less as acceptable (ΔCFI ⩽ .01, 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 (TLI) 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 aforementioned 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 are often misrepresenting 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 $\hat{d}$, 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 (nontau-equivalence) and allows residual item variances to be uncorrelated. SDQ scales are considered sufficiently reliable when ω ⩾ .70, while ⩾.80 is preferred (Evers, Sijtsma, Lucassen, & Meijer, 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

Table 3. Goodness-of-Fit Statistics of the Presumed Five-Factor Structure and the Six-Factor Structure for the SDQ Adolescent Version.

Model	χ^2	df	p	χ^2 Difftest	df Difftest	p	RMSEA	RMSEA 90% [CI]	CFI	ΔCFI	TLI
Five-factor model as hypothesized by Goodman											
Single group											
Clinical	4885.508	265	<.001				.067	[.066, .069]	.850		.831
Community	772.988	265	<.001				.046	[.042, 049]	.896		.883
Multiple group											
Configural inv. I	5451.699	530	<.001				.062	[.061, .064]	.859		.840
Configural inv. II[a]	4271.369	510	<.001				.056	[.054, .057]	.892		.873
Six-factor model (including the positive construal method factor)											
Single group											
Clinical	3862.007	255	<.001				.061	[.059, .062]	.883		.863
Community	525.249	255	<.001				.034	[.030, .038)	.945		.935
Multiple group											
Configural inv. I	4210.048	510	<.001				.055	[.054, .057]	.894		.875
Configural inv. II[b]	4593.298	518	<.001				.053	[.052, .055]	.902		.884
Metric fact. inv.	3879.459	532	<.001	119.060	24	<.001	.051	[.050, .053]	.904	.002	.892
Strong fact. inv.	3852.673	551	<.001	53.286	19	<.001	.050	[.049, .052]	.905	.001	.897
Strict fact. inv.	3901.390	577	<.001	128.589	26	<.001	.049	[.048, .051]	.904	.001	.901

Note. SDQ = Strengths and Difficulties Questionnaire; *df* = degrees of freedom; CI = confidence interval; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; 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 10 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 pair (Q2 and Q10) freed.

exists. We consider Spearman ρ's > .85 to be supportive of the continued use of sum scores in practice.

Results

The SDQ Adolescent Version

Table 3 presents the goodness-of-fit statistics of the single group CFA's in the clinical and community settings, and the table 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 adolescent 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 belonged 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 S1 (Supplementary Material 1; all

Table 4. Unstandardized Parameter Estimates and Standard Errors of the Six-Factor Strict Invariance Model for the SDQ Adolescent Version.

Factor loadings

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

Item	SDQ scale factor loading
Q2-Q10	0.42 (.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

Note. SDQ = Strengths and Difficulties Questionnaire; ES = emotional symptoms; CP = conduct problems; HP = hyperactivity/attention problems; SP = social problems; PB = prosocial behavior; PCM = positive construal method.
*$p > .01$ (for all other values $p < .01$).

Table 5. Goodness-of-Fit Statistics of the Presumed Five-Factor Structure for the SDQ Parent Version.

Five-factor model as hypothesized by Goodman

Model	χ^2	df	p	χ^2 Difftest	df Difftest	p	RMSEA	RMSEA 90% [CI]	CFI	ΔCFI	TLI
Single group											
Clinical	6843.082	265	<.001				.082	[.080, .084]	.848		.828
Community	580.887	265	<.001				.048	[.042, .053]	.926		.916
Multiple group											
Configural inv. I	6785.219	530	<.001				.075	[.073, .076]	.862		.844
Configural inv. II[a]	4972.085	518	<.001				.064	[.062, .065]	.902		.887
Metric fact. inv.	4759.011	538	<.001	62.924	20	<.001	.061	[.059, .063]	.907	.005	.896
Strong fact. inv.	4660.638	558	<.001	74.201	20	<.001	.059	[.057, .061]	.909	.002	.903
Strict fact. inv.	4661.278	589	<.001	199.904	31	<.001	.058	[.056, .059]	.910	.001	.907

Note. SDQ = Strengths and Difficulties Questionnaire; df = degrees of freedom; RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker–Lewis index; 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.

supplementary materials are available in the online version of the article) 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 4.

Adolescents in the community and clinical settings differed from each other regarding their mean psychosocial strengths and difficulties scores: compared with 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 behavior: $\hat{d} = 0.06$, positive construal methods: $\hat{d} = -0.07$).

Adequate reliability was found for the SDQ emotional difficulties, hyperactivity/inattention, and prosocial behavior scales in the clinical and community setting, respectively (emotional difficulties: ω = .85, ω = .81; hyperactivity: ω = .80, ω = .79; prosocial behavior: 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 Version

Table 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 S2 (Supplementary Material 2) shows a representation of the strict invariance model; the factor loadings, residual covariances, factor means, and factor (co)variances are presented in Table 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 6. Compared with 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: $\hat{d} = -1.62$; conduct problems: $\hat{d} = -1.20$; hyperactivity/attention problems: $\hat{d} = -1.41$; social problems: $\hat{d} = -0.88$, and prosocial behavior: $\hat{d} = 0.66$), 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

Table 6. Unstandardized parameter estimates and standard errors of the five-factor strict invariance model for the SDQ parent version.

Factor loadings

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	$\hat{a}$
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

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

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

Table 7. Per SDQ Version and Scale, Cronbach's α and Spearman Rank Correlation Coefficients Between SDQ Scale Scores and Factor Scores.[a]

SDQ scale	SDQ adolescent version		SDQ parent version	
	Six-factor model	Cronbach's α	Five-factor model	Cronbach's α
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

Note. SDQ = Strengths and Difficulties Questionnaire; ES = emotional symptoms; CP = conduct problems; HP = hyperactivity/attention problems; SP = social problems; PB = prosocial behavior.
[a]For all correlation coefficients: $p < .01$.

difficulties: $\omega = .81$, $\omega = .83$; conduct problems: $\omega = .81$, $\omega = .76$; hyperactivity/inattention problems: $\omega = .80$, $\omega = .83$; social problems: $\omega = .77$, $\omega = .82$; prosocial behavior: $\omega = .82$, $\omega = .83$).

Evaluating the Sum Score Method Used in Practice

Table 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 adolescent version ranging from .90 for conduct problems scale to .98 for the hyperactivity/attention problems scale, and for SDQ parent version ranging from .92 for the prosocial behavior scale to .97 for the emotional problems scale. For the sake of comparability with other studies, Table 7 additionally presents Cronbach's alpha coefficient per SDQ scale.

Discussion

This study evaluated the presumed five-factor structure and, if necessary, an alternative factor structure of the SDQ adolescent and the parent versions in clinical and community samples of Dutch adolescents aged 12 to 17 years. 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 Adolescent Version: Factor Structure and Measurement Invariance

For the SDQ adolescent 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 with previous findings. With regard to the community setting our findings are in line with some previous studies (e.g., 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 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 Northeastern 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 adolescent 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 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. Furthermore, evidence was found for measurement invariance of this six-factor structure across clinical and community settings. This finding suggests that the SDQ adolescent 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 Version: Factor Structure and Measurement Invariance

For the SDQ parent 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. Furthermore, evidence was found for measurement invariance of the five-factor structure across clinical and community settings, as was hypothesized. Extending on Smits et al.'s (2016) similar observations regarding children, our findings suggest that the SDQ parent 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/attention problems factor, was found for both SDQ versions (i.e., the five-factor model for the SDQ adolescent version and the six-factor model for the SDQ parent 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 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 attention within the hyperactivity/attention factor (Bøe et al., 2016; van de Looij-Jansen et al., 2011), which is not surprising as the hyperactivity/attention 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/attention factor are strongly associated, and together can be used to sensibly measure hyperactivity/attention.

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 scales 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 problems scales of the SDQ adolescent 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 scales, the clinical setting data show limited variance in scores compared with 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 adolescent 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 contentwise; no corresponding SDQ scale exists. Therefore, only the five existing scales were evaluated for use in practice.

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, as our aim was to evaluate factor structures that closely resemble how the SDQ is used in daily practice.

Our study is the first to assess measurement invariance of the SDQ adolescent and parent 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 in practice. Comparing scores across these settings is, for instance, important for clinicians as they are often interested in how a referred adolescent's scores compared with adolescents from a nonclinical population.

Furthermore, the current study evaluates the factor structure and measurement invariance of multiple SDQ versions, whereas most studies investigate 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 adolescent version. For the parent 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 7 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 7 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. Per SDQ version and for each of the five SDQ scales, we compared sum scores and factor scores. For both SDQ versions, strong association was 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 adolescent version was not evaluated for use in practice, because this is a methodological factor that does not contribute to measurement of psychosocial functioning contentwise. 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.

Our findings further show the conduct and social scales of the SDQ adolescent 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.

Authors' Note

Declaration of Conflicting Interests

Funding

ORCID iDs

Jorien Vugteveen (iD) https://orcid.org/0000-0002-8098-4120
Annelies de Bildt (iD) https://orcid.org/0000-0002-4196-2404

Supplemental Material

Supplemental material for this article is available online.

References

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.

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

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*, e0152202.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

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*, 396-406.

Evers, A., 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*, 295-317.

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. doi:10.1186/1744-859X-8-20

Ginkel, J. R., 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*, 315-337.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of child Psychology and Psychiatry*, *38*, 581-586.

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

He, J., Burstein, M., Schmitz, A., & Merikangas, K. R. (2013). The Strengths and Difficulties Questionnaire (SDQ): The factor structure and scale validation in U.S. adolescents. *Journal of Abnormal Child Psychology*, *41*, 583-595.

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*, 1-55.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*, 29-51.

Koskelainen, M., Sourander, A., & Vauras, M. (2001). Self-reported strengths and difficulties in a community sample of Finnish adolescents. *European Child & Adolescent Psychiatry*, *10*, 180-185.

Lundh, L.-G., 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*, 523-532.

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479-515.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

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.

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*, 603-610.

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 multiethnic urban adolescent sample. *Journal of Child Psychology and Psychiatry*, *52*, 1002-1011.

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*, 73-82.

Ruchkin, V., Koposov, R., & Schwab-Stone, M. (2007). The Strength and Difficulties Questionnaire: Scale validation with Russian adolescents. *Journal of Clinical Psychology*, *63*, 861-869.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity*, *24*, 367-386.

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

Smits, I. A., Theunissen, M. H., 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*, 238-246.

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

Steiger, J. H. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the Annual Spring Meeting of the Psychometric Society, IA.

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*, 127-144.

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*, 1304-1312.

Van Widenfelt, B. M., Goedhart, A. W., Treffers, P. D., & Goodman, R. (2003). Dutch version of the Strengths and Difficulties Questionnaire (SDQ). *European Child & Adolescent Psychiatry*, *12*, 281-289.

Youngstrom, E. A. (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*, 139-159.

Youngstrom, E. A., & Frazier, T. W. (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: Wiley.